UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Análise por Prospeção de Dados de Registos Eletrónicos de Saúde de Pacientes com Cancro do Pulmão

Mestrado em Gestão de Informação

Gestão e Análise de Dados

Ana Cristina Antunes da Silva

Trabalho de Projeto orientado por:

Prof.ª Doutora Cátia Luísa Santana Calisto Pesquita

Prof.ª Doutora Lisete Maria Ribeiro de Sousa

2016

O que hoje não sabemos,

amanhã saberemos

(Garcia de Orta, 1563).

Agradecimentos

Em primeiro lugar quero agradecer às minhas orientadoras Prof.ª Cátia Pesquita e Prof.ª Lisete Sousa pelo interesse, disponibilidade e críticas ao longo destes últimos meses que me ajudaram a crescer. Também quero agradecer a confiança que depositaram em mim no decorrer deste estudo.

Quero também agradecer à Dra. Ana Miranda e Dra. Alexandra Mayer pela oportunidade de me terem facultado os recursos do Registo Oncológico Regional Sul. O meu muito obrigado a toda a equipa por me terem recebido tão bem e por todos os comentários e críticas que foram fazendo ao longo destes meses.

Aos meus colegas de mestrado e em particular à Sara, Inês e Diogo, quero agradecer o companheirismo e os bons momentos que proporcionaram durante o curso.

Por fim mas não menos importante, quero agradecer aos meus pais e ao meu irmão por me terem dado ao longo da vida a liberdade para saltar mais alto do que seria possível e pelo amor incondicional que me ajudou a ultrapassar todas as barreiras. Ainda aos meus amigos, e em particular ao Tiago, Joana, Sandra, Rebeca, Isabela e Mafalda, que também me ajudaram a crescer e tornaram a minha vida mais valiosa.

Mais uma vez, a todos o meu muito obrigado. Espero que este trabalho vos deixe tão orgulhosos de mim como eu estou dele.

Resumo

As informações de saúde dos indivíduos ao longo da sua vida, por exemplo a partir das anotações médicas, são registadas em sistemas de bases de dados, dando origem aos registos eletrónicos de saúde. Devido ao crescente interesse pela recolha, registo e armazenamento destes dados, o seu volume atinge atualmente proporções elevadas e é cada vez mais importante e valorizado conseguir usar esta informação como ferramenta para que as entidades de saúde adquiram maior conhecimento acerca das doenças.

Hoje em dia utiliza-se frequentemente técnicas de estatística clássicas mas devido ao volume de informação estas técnicas tornam-se insuficientes ou até obsoletas. As técnicas de prospeção de dados vieram colmatar esta insuficiência, através da sua automatização e eficiência computacional.

Este estudo tem como objetivo a extração de conhecimento útil a partir dos registos eletrónicos de saúde dos pacientes com cancro do pulmão. Estes pacientes foram diagnosticados no decorrer do primeiro semestre de 2013 e residem na região abrangida pelo Registo Oncológico Regional Sul.

Paralelamente ao objetivo principal deste estudo, pretende-se integrar registos eletrónicos de saúde dos pacientes oncológicos com dados relativos a comportamentos de risco individuais e a fatores ambientais, uma vez que diversos estudos referem a existência de uma relação entre esses fatores. Pretende-se, ainda, descobrir padrões da incidência do cancro do pulmão, a nível geográfico, na região sul de Portugal.

Em primeiro lugar, aplicaram-se técnicas de estatística descritiva e de inferência estatística para se conhecer a estrutura e as características do conjunto de dados recolhidos. Posteriormente, para estudar padrões de incidência do cancro do pulmão a nível geográfico, aplicaram-se técnicas de autocorrelação e associação espacial. Por fim, aplicaram-se métodos de agrupamento - nomeadamente de agrupamento hierárquico e de particionamento - utilizando como referência as circunstâncias demográficas, características do tumor, comportamentos de risco e fatores ambientais, com vista a encontrar grupos de pacientes com caraterísticas semelhantes entre si. Os algoritmos e métodos de agrupamento explorados foram avaliados por medidas de qualidade por forma a obter o melhor particionamento dos dados.

Em cada análise realizada, pretendeu-se construir um modelo descritivo adequado a qualquer conjunto de dados (com características semelhantes ao do conjunto de dados em

análise), com o objetivo de, numa forma rápida e automática, encontrar relações e padrões subjacentes ao conjunto de dados que leve à obtenção de conhecimento útil.

Relativamente aos resultados finais, verificou-se que a utilização e combinação de diversas técnicas complementares proporcionam uma maior segurança e confiança nos resultados obtidos; os modelos construídos podem ser aplicados a outros conjuntos de dados com características semelhantes, facultando uma análise eficiente de um grande conjunto de dados em curto espaço de tempo. Concluiu-se que os fatores ambientais e a idade média dos pacientes por concelho têm um impacto direto na taxa de incidência do cancro do pulmão e que existem diferenças significativas a nível geográfico que carecem de uma investigação mais profunda.

A principal conclusão deste estudo é de que uma análise mais abrangente dos registos eletrónicos de saúde dos pacientes oncológicos pode permite encontrar relações significativas entre alguns fatores presentes no estudo. Foi ainda revelada a importância da abrangência e completude dos dados para o sucesso deste tipo de investigação. Estes resultados poderão abrir portas a novas linhas de estudo e ao estabelecimento de objetivos mais concretos em futuras investigações.

Palavras-chave: Análise de Agrupamentos, Análise de Dados Espaciais, Cancro, Registos Eletrónicos de Saúde, Prospeção de Dados.

Abstract

During their lifetime, people's health information (such as notes of medical records) is stored in database systems, yielding electronic health records. Currently the volume of such data reaches very high proportions due to the growing interest in collecting, recording and storing them. Therefore, the ability to use these data as a tool has a greater importance for the health authority to acquire new knowledge of numerous diseases. Traditional statistical techniques are still commonly used today. However, due to the large volume of data, these techniques end up being insufficient or even obsolete. Data Mining has filled that gap because it presents technical features such as automation and computational efficiency.

The goal of this research is to extract knowledge from the electronic health records of lung cancer patients using mainly data mining techniques. These patients were diagnosed during the first semester of 2013 and reside in the area covered by the Registo Oncológico Regional Sul. Another purpose of this research is integrating the health information of cancer patients with data concerning individual risk behaviors and environmental factors because there are several studies reported the existence of a relationship between the above criteria. Finally, we intended to conduct a geographical analysis of lung cancer incidence in southern Portugal, in order to find patterns of incidence geographically.

In this study, first we applied techniques of descriptive statistics and statistical inference to know the structure and the characteristics of dataset collected. The next step, in order to study the incidence of geographically lung cancer, was to apply autocorrelation and spatial association techniques. Then, we applied clustering methods, including hierarchical clustering methods and partitioning, in order to find groups of patients with similar characteristics to each other, using as reference the demographic features, tumor characteristics, risk behaviors and environmental factors. It is intended associating quality measures in addition to the same algorithms and clustering methods in order to explore and try different strategies to achieve better results. In both analyzes intended to build an adequate descriptive model for any dataset, in which the goal is to find hidden patterns and relationships in the dataset that leads to obtaining useful information, a process that is both fast and automated.

The results suggest that the use and combine several complementary techniques provide a better quality for research and also greater confidence in conclusions. Another contribution from this work are the constructed models that can be applied to other datasets with similar characteristics and can provide an efficient analysis of a large dataset in real time.

It was also found that environmental factors and the average age of a certain county patients have a direct impact on the incidence rate of lung cancer. At last, significant differences geographically were found, which need more research in the near future.

The main conclusion is that the application of this type of methods allows us to draw a more comprehensive analysis of the electronic health records of cancer patients, which can support the finding of some significant relationships between certain factors considered in the study. Moreover, the importance of a broad scope and completeness of data was identified. These results could open doors to new lines of research and the establishment of more concrete goals in studies to be undertaken in the future.

Keywords: Clustering, Spatial Data Analysis, Cancer, Electronic Health Records, Data Mining.

Índice

Índice de Figuras

Índice de Tabelas

Siglas e Acrónimos

APA	Agência Portuguesa do Ambiente
DGT	Direção-Geral do Território
INE	Instituto Nacional de Estatística
INS	Inquérito Nacional de Saúde
IQA	Índice de Qualidade do Ar
LVT	Lisboa e Vale do Tejo
PD	Prospeção de Dados
RAM	Região Autónoma da Madeira
RES	Registos Eletrónicos de Saúde
ROR	Registos Oncológicos Regionais
ROR-Sul	Registo Oncológico Regional Sul

1 INTRODUÇÃO

Este documento descreve um estudo focado na análise exploratória de registos clínicos mantidos pelo Registo Oncológico Regional Sul (ROR-Sul), inserido no Instituto Português de Oncologia de Lisboa Francisco Gentil. Neste capítulo apresentam-se as motivações e objetivos deste estudo, o plano de trabalho desenvolvido, uma breve descrição da estrutura e missão do ROR-Sul e, por fim, a organização deste documento.

1.1 MOTIVAÇÃO

Os Registos Eletrónicos de Saúde (RES) constituem uma fonte de dados diversificada que agrega as informações de saúde dos indivíduos ao longo da sua vida. Apesar do seu potencial, a sua exploração está ainda aquém do desejável, devido essencialmente à complexidade da estrutura dos dados clínicos e à grande quantidade de valores omissos que constituem impedimentos na utilização científica dos RES. Uma das vertentes de investigação é a prospeção dos RES, tendo como principais objetivos: descobrir novas estratificações dos pacientes; investigar as características dos pacientes, do diagnóstico e dos tratamentos; e identificar possíveis correlações entre diferentes doenças que até ao momento eram desconhecidas. A prospeção dos RES conjuntamente com outros dados relevantes pode, além disso, contribuir beneficamente na tomada de decisões clínicas (Iakovidis, 1998; Roque et al., 2011; Hagar et al., 2014; Jensen, Jensen e Brunak, 2012).

A Prospeção de Dados (PD) é um conjunto de métodos e técnicas computacionais que permite descobrir padrões e relações relevantes e inesperadas a partir de uma grande quantidade de dados de natureza diversa, constituindo-se como um tópico de investigação bastante vasto e em expansão ativa. A sua natureza não saturada deve-se principalmente ao facto de ir ao encontro de uma necessidade da sociedade e de se enquadrar numa vertente tecnológica em constante evolução. A capacidade da PD para lidar com grandes conjuntos de dados diversos torna-a numa estratégia particularmente indicada para a investigação de fenómenos clínicos complexos, como por exemplo, as doenças oncológicas (Fayyad, Piatetsky-Shapiro e Smyth, 1996; Hand, Mannila e Smyth, 2001; Han, Kamber e Pei, 2011).

Perante o atual impacto na sociedade, a nível mundial e nacional (Parente et al., 2007; Registo Oncológico Nacional [RON] 2006, 2012), tem-se verificado um crescente interesse na

investigação na área do cancro, sobretudo através da prospeção dos RES - uma via mais económica e acessível em tempo real (Hagar et al., 2014).

As doenças oncológicas afetam uma parte considerável da população mundial. Calcula-se que em 2012, a nível mundial, tenham ocorrido cerca de 14 milhões de novos casos de cancro em 7 mil milhões de habitantes (com exceção dos casos de cancro da pele do tipo não melanoma). Em Portugal, estes números ascendem aproximadamente a 49,2 mil novos casos de cancro em 10,7 milhões de habitantes (International Agency for Research on Cancer [IARC], 2015) e estima-se que nas próximas duas décadas o número de novos casos em todo o mundo tenha um crescimento de 70% (World Health Organization [WHO], 2015)

Na Figura 1.1 são apresentadas as taxas mundiais (incluindo as portuguesas) de incidência e mortalidade causadas pelo cancro, em 2012, por 100 000 habitantes, padronizadas pela idade: de um modo geral, verifica-se em Portugal que ambas as taxas são elevadas, reforçando assim a importância de estudos sobre o cancro em Portugal.

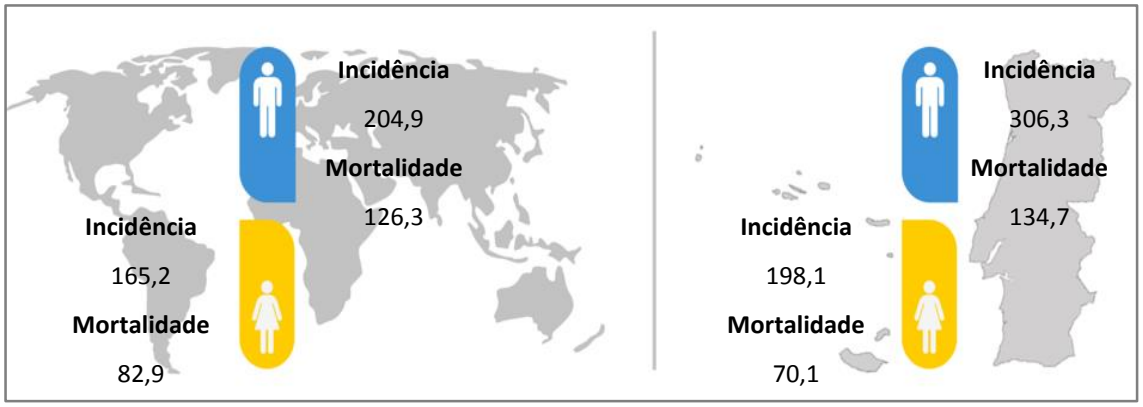

Figura 1.1 - Taxas e incidência e mortalidade padronizadas pela idade, por 100 000 habitantes, provocadas pelo cancro, exceto cancro da pele do tipo não melanoma, por sexo, em 2012 no Mundo e em Portugal (adaptado de IARC, 2015)

Segundo os registos de 2012, de entre os diversos tipos de cancro existentes, o que revelou maior mortalidade foi o cancro do pulmão, a nível nacional e mundial.

Quando observamos os tipos de cancro com maior incidência e mortalidade por sexo existem algumas diferenças (IARC, 2015), nomeadamente:

- no sexo feminino, em termos da incidência e mortalidade, predomina o cancro da mama, em Portugal e a nível mundial;

- no sexo masculino, a nível mundial o cancro do pulmão tem a maior incidência, enquanto em Portugal é o cancro da próstata; em relação à mortalidade, tanto a nível nacional como mundial, é o cancro do pulmão.

O foco deste trabalho reside na análise dos registos clínicos de pacientes diagnosticados com cancro do pulmão, tendo sido o tipo de cancro sugerido por especialistas do ROR-Sul devido à sua elevada incidência e mortalidade em Portugal e em indivíduos de ambos os sexos.

1.2 OBJETIVOS

O principal objetivo deste estudo é a obtenção de informação útil contida nos RES dos pacientes com cancro do pulmão (residentes nas áreas geográficas abrangidas pelo ROR-Sul), com o intuito de reforçar e completar o perfil das características destes pacientes.

Para atingir este objetivo foram considerados três desafios complementares:

- extração, limpeza e integração de dados;
- análise geográfica da incidência do cancro do pulmão;
- descoberta de padrões no conjunto de dados composto pelos RES e pelos dados externos.

Na primeira vertente, para além da seleção de variáveis e limpeza dos dados, serão também integrados dados, como por exemplo hábitos tabágicos ou poluição atmosférica, provenientes de outras fontes externas. Com esta inclusão pretende-se investigar os desafios ao nível da integração de fontes de dados públicas com os registos clínicos. Pretende-se ainda possibilitar uma análise conjunta dos dados de forma a identificar e encontrar possíveis relações entre os comportamentos de risco dos indivíduos e os fatores ambientais, que caracterizam os contextos de vida dos pacientes com cancro do pulmão da região ROR-Sul.

Na segunda vertente pretende-se analisar a incidência do cancro do pulmão ao nível geográfico na região ROR-Sul, com o propósito de investigar a utilidade de métodos de análise de dados espaciais na identificação de padrões úteis, que possam auxiliar a compreensão do fenómeno do cancro a nível populacional.

Na terceira vertente pretende-se não só descobrir padrões no conjunto de dados em estudo através da análise de agrupamentos, como investigar a aplicação de diferentes medidas de distância entre observações e o seu agrupamento, tendo como objetivo a deteção de grupos de pacientes com características comuns para suportar uma possível estratificação a diferentes níveis, como por exemplo, clínico, geográfico, ambiental, etc..

1.3 PLANO DE TRABALHO

A metodologia estabelecida para atingir estes objetivos incluiu a realização de tarefas, de base, de exploração e tratamento dos dados, nomeadamente:

1. escolha do *software* de análise de dados;
2. definição da amostra;
3. seleção das variáveis de interesse;
4. extração dos RES presentes no ROR-Sul;
5. extração dos dados de outras fontes de dados;
6. pré-processamento dos dados recolhidos;
7. integração de dados recolhidos;
8. estatísticas descritivas a partir dos dados;
9. inferência estatística - testes de hipóteses.

A realização destas tarefas permitiu construir um suporte para realização das seguintes análises:

→ análise de dados espaciais;

→ análise de agrupamentos.

Estas tarefas e análises irão permitir que os objetivos mencionados anteriormente sejam atingidos.

A primeira análise foca-se na distribuição geográfica da taxa de novos casos (taxa de incidência) com cancro do pulmão na mesma região, através da aplicação de métodos de autocorrelação e associação espacial, que permitem obter grupos de regiões segundo a sua semelhança (ou dissemelhança) com as regiões vizinhas.

A segunda análise foca-se na caracterização mais detalhada dos pacientes com cancro do pulmão da região ROR-Sul, tendo sido consideradas tanto as características demográficas e dos tumores dos pacientes, como também os comportamentos e fatores ambientais relacionados com a ocorrência do cancro do pulmão. Pretendeu-se aplicar métodos de agrupamento hierárquico aglomerativo e de particionamento dos dados, com a finalidade de combinar as vantagens de ambos os métodos. Com o objetivo de encontrar o melhor agrupamento possível dos dados, tentou-se comparar diversos métodos de agrupamento hierárquico (como por exemplo: método de *ward* e o método da média do grupo), utilizando medidas de qualidade dos agrupamentos obtidos para permitir uma primeira avaliação dos mesmos. Relativamente aos métodos de particionamento, pretende-se também comparar os

diversos resultados obtidos através, por exemplo, da silhueta de cada observação em cada grupo encontrado.

Destaca-se que, embora as duas análises referidas se processem de forma independente e distinta, elas são complementares. Para além dos resultados contribuírem para uma coesão da informação obtida, as duas análises contribuem igualmente para um melhor conhecimento do cancro do pulmão e das características dos seus pacientes na região sul de Portugal.

Por fim, pretende-se em cada uma das análises construir um modelo descritivo adequado a qualquer conjunto de dados, mantendo uma estrutura específica, de modo a que possa ser replicado para a investigação de outras doenças.

1.4 Registo Oncológico Regional Sul

O ROR-Sul é um dos quatro Registos Oncológicos Regionais (ROR) que no seu conjunto agregam os registos oncológicos a nível nacional (RON 2006, 2012). Assim, será revelante compreender a importância e a evolução do registo oncológico em Portugal de forma a enquadrar o atual papel do ROR-Sul.

Em 1988, com a Portaria nº 35/88 de 16 de Janeiro, o registo do cancro passa a ser obrigatório em Portugal, tendo sido criados três ROR em cada um dos centros regionais do Instituto Português de Oncologia de Francisco Gentil. Nesta Portaria ficou também determinado que cada hospital (central ou distrital) teria de criar um registo oncológico, e sendo que estas entidades ficariam obrigadas a remeter os registos recolhidos ao ROR da sua área geográfica (Lunet e Pimentel, 2012).

Para cumprir esta lei tem-se procurado, ao longo dos anos, combinar esforços para cultivar a necessidade do registo oncológico e o seu processamento para o apoio da decisão médica. O registo oncológico é a chave para avaliar e controlar o impacto das doenças oncológicas numa dada região, tornando-se indispensável para o planeamento, monitorização e avaliação das campanhas de prevenção das doenças oncológicas (RON 2006, 2012).

Atualmente em Portugal existem quatro ROR, como se pode verificar na Tabela 1.1. Nestes ROR são registadas manualmente as informações de saúde dos pacientes oncológicos de todas as instituições públicas de saúde (tais como os hospitais e centros de saúde), de todos os distritos de Portugal continental e das regiões autónomas, e ainda de algumas instituições privadas, como os hospitais e laboratórios (RON 2006, 2012). Há alguns anos que os ROR cobrem na sua totalidade o território nacional, encontrando-se Portugal numa situação privilegiada em

relação aos restantes países europeus no que toca à cobertura e antiguidade dos registos oncológicos (Lunet e Pimentel, 2012).

Tabela 1.1 - Descrição dos distritos e regiões autónomas pertencentes a cada ROR

ROR	Distritos e Regiões Autónomas
RORRENO	Braga, Bragança, Porto, Viana do Castelo, Vila Real
RORCENTRO	Aveiro, Viseu, Guarda, Coimbra, Castelo Branco, Leiria
ROR-Sul	Lisboa, Santarém, Setúbal, Portalegre, Évora, Beja, Faro, Leiria e Região Autónoma da Madeira
RORA	Região Autónoma dos Açores

Nota: adaptado de RON 2006, 2012, p. 11

Em particular, o ROR-Sul, sedeado nas instalações do IPOLFG, tem como principal foco a criação de uma base de dados que sirva como ferramenta para a monitorização do cancro na sua área de abrangência. Analogamente aos restantes registos oncológicos regionais, o ROR-Sul tem como missão a recolha, tratamento e processamento dos dados referentes aos tumores malignos que ocorram na população residente na área abrangida pelos seus serviços (ROR-Sul, s.d.).

A Figura 1.2 ilustra o território nacional (Direção Geral do Território [DGT], 2015) abrangido pela região ROR-Sul (consultar o Anexo 1 para examinar com mais pormenor a região), que se divide em quatro regiões: Lisboa e Vale do Tejo (LVT), Alentejo, Algarve e Região Autónoma da Madeira (RAM). Estas regiões são agregações dos distritos apresentados na Tabela 1.1 e na Figura 1.2, destacando-se que o distrito de Leiria está dividido em dois ROR, como indicado na Tabela 1.1. Ao ROR-Sul pertencem os seguintes concelhos: Alcobaça, Bombarral, Caldas da Rainha, Nazaré, Óbidos e Peniche.

Salienta-se que o ROR-Sul é um dos maiores registos oncológicos regionais da Europa, uma vez que abrange uma área com uma ponderação próxima de 50% do território nacional e uma população média anual próxima de 4,8 milhões de habitantes (ROR-Sul, 2014).

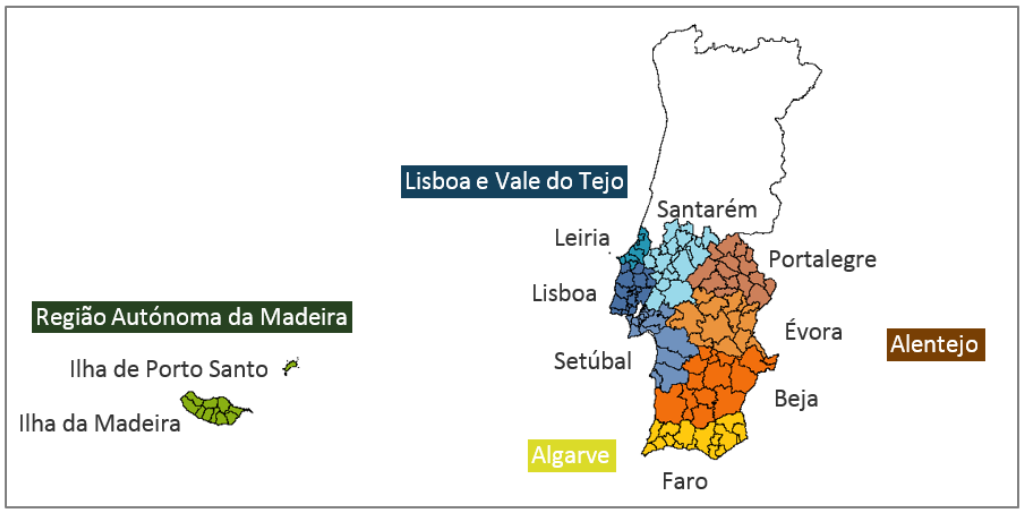

Figura 1.2 - Área territorial de Portugal abrangida pelo ROR-Sul (adaptado de ROR-Sul, 2014; DGT, 2015)

1.5 Organização do Documento

Este documento descreve todos os desafios encontrados, as soluções aplicadas e os resultados obtidos no decurso deste estudo. Tendo sido organizado através de uma sequência lógica que permite ao leitor compreender o seguimento a sequência das tarefas realizadas e estratégias escolhidas ao longo do estudo desenvolvido.

O presente documento segue as normas sugeridas na sexta edição do *Publication Manual of the American Psychological Association* (American Psychological Association, 2012).

De forma a enquadrar teoricamente o tema, no próximo capítulo serão abordados os seguintes conceitos: registos eletrónicos de saúde, prospeção de dados, análise de dados espaciais, análise de agrupamentos, cancro e cancro do pulmão.

No capítulo 3 será descrita a metodologia, nomeadamente as tarefas e estratégias utilizadas que marcaram o caminho deste estudo, em concordância com a informação analisada na literatura. No capítulo 4 serão apresentados os resultados obtidos. Os resultados serão discutidos no capítulo 5 e, por fim, as conclusões gerais deste estudo serão descritas no capítulo 6.

2 REVISÃO DA LITERATURA

Este capítulo introduz alguns conceitos teóricos e essenciais para a compreensão da temática deste estudo, nomeadamente: RES; PD; análise de dados espaciais; análise de agrupamentos; e o cancro. São também descritos trabalhos de investigação recentes relacionados com o tema deste trabalho.

2.1 REGISTOS ELETRÓNICOS DE SAÚDE

Os RES são uma ferramenta indispensável para o aperfeiçoamento dos cuidados de saúde continuados. Estes surgiram com base nas notações médicas tiradas no contexto da prática clínica prestada pelos profissionais de saúde aos pacientes (Iakovidis, 1998; Anderson, 2007). Será por isso relevante contextualizar a evolução do processo de registo de notações médicas que deu origem aos RES.

Há aproximadamente cinco décadas, desenvolveu-se o interesse pela uniformização das notações médicas, processo que consistiria, sobretudo, na implementação de normas de registo por parte dos profissionais de saúde. Esta uniformização iria, não só facilitar a transmissão de informação entre os profissionais de saúde, como também permitir que estas notações fossem recolhidas, registadas e armazenadas em sistemas de bases de dados, dando origem aos RES, prevendo-se a utilidade futura destes registos (Weed, 1968).

Atualmente existem sistemas de bases de dados adaptados ao nível de complexidade dos RES, aumentando assim o interesse pelo registo das informações de saúde: estes registos podem ser utilizados em estudos estatísticos sobre grupos de pacientes, na deteção de problemas nos cuidados médicos individuais, no apoio da continuidade dos cuidados de saúde e educação e ainda, assegurar a confidencialidade de forma intemporal (Weed, 1968; Iakovidis, 1998; Anderson, 2007; Hagar et al., 2014).

A aplicação dos RES, em particular, à investigação científica pode permitir obter mais informação acerca de determinada doença ou estudar o processo dos cuidados de saúde ao longo dos anos. Contudo, neste tipo de estudo podem existir alguns obstáculos: o facto de a classificação dos termos clínicos sofrer alterações ao longo do tempo, dando origem a dados inconsistentes; ou, dada a mobilidade dos pacientes, o facto de recorrerem a cuidados de saúde

em diferentes instituições levanta o problema de o histórico clínico de um paciente, numa certa instituição, não estar completo.

Ainda no âmbito anterior, um conjunto de dados constituído por dados inconsistentes ou omissos pode traduzir-se, por exemplo, numa fraca aplicação de métodos estatísticos. É de salientar ainda, que a ausência de registo não significa necessariamente que o indivíduo goze de boa saúde, mas sim que as ocorrências ao longo do tempo não tenham sido registadas na mesma instituição.

Existem várias formas de prevenir este acontecimento: alargar a área geográfica em estudo de modo a incluir várias instituições de saúde; escolher instituições - como os hospitais distritais - que agreguem um maior número de pacientes; ou ainda, selecionar uma subpopulação constituída por pacientes que residam há muito tempo na mesma região (Hagar et al., 2014).

Apesar do potencial dos RES, o seu processamento continua aquém do esperado ao contrário do que se verifica noutros setores económicos. Nos tópicos seguintes são apresentados alguns desafios à partilha dos sistemas que agregam os RES (Iakovidis, 1998; Anderson, 2007).

- **Questões organizacionais e culturais -** algumas culturas não abraçam a ideia de partilha de informações médicas dos pacientes, sendo o profissional de saúde penalizado por partilhar informações com os seus colegas. Geralmente o sistema de RES é privado e independente, não havendo incentivos para a partilha de informação devido, principalmente, à competição entre os centros de cuidados de saúde. No cerne deste problema está a questão de saber se os dados são propriedade do médico, do paciente ou da entidade de saúde.
- **Tecnologia -** a este nível, os maiores desafios residem na utilização de diferentes plataformas tecnológicas e sistemas de base de dados heterogéneos para armazenamento, manutenção, comunicação e recuperação de informação multimédia (e.g.: texto; resultados de testes; imagens; pagamentos), tornando bastante morosa a integração dos diversos sistemas (e.g.: administração; seguro; clínico; enfermagem).
- **Uniformidade -** o processo de uniformização de todos os sistemas, classificação e codificação de doenças, e procedimentos médicos, tem sido um processo gradual, onde a demora dos procedimentos e a falta de financiamento são os principais obstáculos à adoção de critérios internacionais uniformes.

Há ainda outros aspetos que dificultam a uniformização:

- a dificuldade em desenhar uma arquitetura homogénea do sistema que possa agregar todo o tipo de dados preservando a forma original do seu conteúdo e contexto. Além de serem despendidos muitos recursos a tentar resolver este problema, cada prestador de cuidados de saúde tem necessidades funcionais diferentes e, portanto, consideram mais viável desenhar uma estrutura de dados para o sistema que vá ao encontro das mesmas, em detrimento da homogeneidade das estruturas de dados em todos os sistemas;
- a complexidade de partilha, integração ou troca de informações com outras fontes de conhecimento, devido ao facto de que cada sistema geralmente seguir uma codificação ou classificação das doenças e procedimentos diferente.

Em suma, os RES contêm inúmeras informações sobre os cuidados médicos registados ao longo do tempo sobre cada paciente, proporcionando uma fonte rica em informação que pode ser usada, por exemplo, na investigação científica. Este tipo de estudo, extração de conhecimento a partir dos dados, é ainda uma vantagem em comparação com os tradicionais estudos em epidemiologia, devido a uma menor complexidade no processo de extração de conhecimento de dados (Hagar et al., 2014). Assim, reforça-se a importância do registo sistemático e eletrónico das notações médicas para a melhoria dos cuidados de saúde em geral.

2.2 Prospeção de Dados

A PD foca-se no processo de avaliar a melhor forma de usar os dados para descobrir regularidades globais e melhorar o processo de decisão (Mitchell, 1999). Deste modo, importa, em primeiro lugar, apresentar a história da prospeção de dados para compreender a sua origem e a sua importância na atualidade.

Com o avanço da tecnologia, o processo de registo e armazenamento de dados tornou-se automático dando origem a grandes bases de dados em suporte eletrónico, constituídas por informações úteis para tomada de decisões. No entanto, a dificuldade em gerir, tratar e analisar estes dados, leva a que geralmente as decisões sejam tomadas com base na intuição dos decisores, em vez de no conhecimento extraído dos dados.

Posteriormente, a redução do custo de armazenamento destes grandes conjuntos de dados, o aumento da facilidade de registo de dados, e o desenvolvimento de algoritmos de, por

exemplo, aprendizagem automática sólidos e capazes de processar eficientemente os dados, despertaram o interesse em explorar o conteúdo das bases de dados, transformando-o em conhecimento organizado e útil (Fayyad et al., 1996; Mitchell, 1999).

Assim, a PD é o processo de pesquisa por padrões interessantes e relações inesperadas em grandes conjuntos de dados, e a transferência destes resultados para um formato compreensível e informativo para o decisor.

Por vezes o termo PD é considerado como sinónimo de Extração de Conhecimento a partir dos Dados (termo traduzido do original em língua inglesa, *Knowledge Discovery from Data*). No entanto, numa outra perspetiva, considera-se que a PD é um dos diversos passos que constituem o processo de Extração de Conhecimento a partir dos Dados, como ilustra a Figura 2.1 (Fayyad et al., 1996; Mitchell, 1999; Hand et al., 2001; Han et al., 2011).

Figura 2.1 - Etapas do processo de Extração de Conhecimento a partir dos Dados (adaptado de Han et al., 2011, pp. 6-8

Dependendo do objetivo da investigação o investigador tem à sua disposição diferentes tipos de tarefas de PD. Seguidamente, apresentam-se algumas das principais tarefas de PD (Fayyad et al., 1996; Hand et al., 2001; Han et al., 2011).

- **Modelos Descritivos -** estes modelos são usados geralmente para caracterizar propriedades ou descrever todo o conjunto de dados num formato mais conveniente, sumário e compreensível. Um exemplo desta análise é a partição dos dados em grupos, sendo designada por análise de agrupamentos ou

segmentação. Esta análise tem como objetivo agrupar observações que sejam semelhantes dentro de cada grupo, e dissemelhantes fora destes. A partição dos dados pode ser utilizada por conveniência prática ou, em contraste, ser utilizada para identificar grupos naturais tendo em conta os dados. Em termos práticos pode-se aplicar modelos descritivos, e em particular a análise de agrupamentos, ao estudo de doenças psiquiátricas com o intuito de identificar as causas destas segundo os agrupamentos constituídos por diferentes perfis de sintomas.

- **Modelos Preditivos -** este tipo de análise baseia-se nos dados para fazer predições. O objetivo consiste em construir um modelo que, através de um conjunto de variáveis, permita predizer o valor de uma outra variável. Dentro dos modelos preditivos, os de regressão e de classificação são os mais utilizados. Uma aplicação prática das técnicas de regressão consiste na construção de modelos que permitem estimar probabilidades, por exemplo: estimar a probabilidade de um número de telefone pretender mudar de operadora, através de um certo número de variáveis explicativas.
- **Descoberta de Padrões e Regras -** permite identificar padrões que ocorram com frequência nos dados. A exploração destes padrões frequentes leva à descoberta de associações e correlações interessantes dentro dos dados. A deteção de padrões pode ser usada para, por exemplo, identificar comportamento fraudulento através da deteção de regiões no espaço definidas por diferentes tipos de operações, onde certos pontos são significativamente diferentes dos restantes. Outra tarefa, bastante célebre, consiste na procura de combinações de observações diferentes que ocorram com uma certa frequência no conjunto de dados. Esta tarefa tem sido muito usada em conjuntos de dados compostos por transações de supermercado para, através de algoritmos baseados em regras de associação, encontrar produtos de supermercado que sejam comprados em simultâneo com uma certa frequência pré-definida pelo investigador.

Todas estas tarefas apresentadas têm em comum:

→ a determinação de um modelo;

→ a forma de avaliar a qualidade do modelo;

→ otimização da avaliação da qualidade dos modelos encontrados;

→ estratégia de transferência da informação extraída para um formato compreensível.

Apesar da importância da adequabilidade do modelo aos dados deve-se evitar modelos demasiado ajustados, pois, neste tipo de análise utiliza-se apenas um subconjunto de dados, e se houver um grande ajustamento do modelo a esse subconjunto de dados, o mesmo não se adaptará ao restante conjunto. Assim, considera-se melhor dar preferência a modelos que generalizem o subconjunto de dados (Fayyad et al., 1996; Hand et al., 2001).

A PD adota diversas técnicas de outros domínios, como ilustrado na Figura 2.2, onde a natureza multidisciplinar da PD e a diversidade de áreas de aplicação contribui claramente para o seu sucesso. Todavia, do mesmo modo que é difícil definir fronteiras entre estes domínios, é também difícil definir fronteiras entre cada um destes e a PD (Fayyad et al., 1996; Hand et al., 2001; Mitchell, 1999; Han et al., 2011).

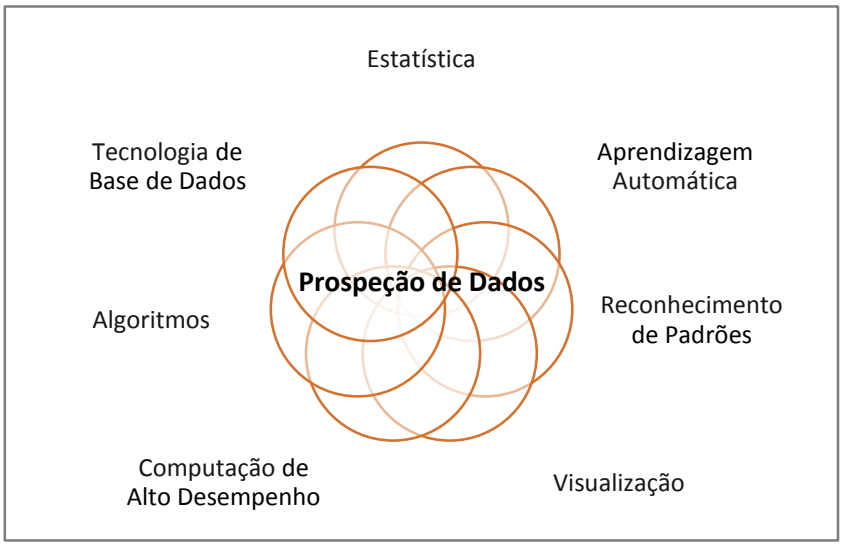

Figura 2.2 - Domínios da PD (adaptado de Han et al., 2011, p. 23)

No caso particular da estatística, as suas técnicas quando aplicadas isoladamente podem não ser suficientes para fazer face a alguns desafios que o processamento de grandes conjuntos de dados acarreta. Desta forma, a diferença fundamental entre a aplicação de métodos estatísticos clássicos e de PD é a dimensão do conjunto de dados sobre o qual recai. Quando se explora grandes quantidades de informação, através dos primeiros métodos indicados,

características evidentes poderão escapar ao investigador, uma vez que, neste caso não basta apenas uma simples visualização. Deste modo, pesquisas mais sofisticadas como a PD, poderão reconhecer em grandes conjuntos de dados, características que seriam evidentes em menores quantidades de informação.

No entanto, a estatística desempenha um papel relevante, tornando-se uma componente necessária da PD: os modelos estatísticos descrevem o comportamento das observações e podem ser usados para sumariar ou descrever um conjunto de dados, além da sua utilidade para inferir a população em estudo, por exemplo, no que toca à aleatoriedade das observações (Fayyad et al., 1996; Hand et al., 2001; Han et al., 2011).

Outro domínio que requer atenção é o de aprendizagem automática. Esta ciência investiga o modo como os computadores podem aprender ou melhorar o seu desempenho com base nos dados. A maior investigação desta ciência centra-se nos programas computacionais que aprendem de forma autónoma a reconhecer padrões complexos e a tomar decisões inteligentes baseando-se nos dados.

Existem duas abordagens de aprendizagem automática relacionadas com a PD: aprendizagem supervisionada e aprendizagem não supervisionada. A primeira abordagem é análoga à tarefa de classificação - caracteriza-se por existir à partida uma classe para cada observação do conjunto de dados que supervisiona a aprendizagem do modelo de classificação. A segunda abordagem é análoga à tarefa de agrupamento - o processo de aprendizagem não supervisionada tem esta designação porque as observações do conjunto de treino não têm uma classe atribuída à partida, recorrendo-se geralmente aos algoritmos de agrupamento para descobrir possíveis classes dos dados (Fayyad et al., 1996; Han et al., 2011).

2.2.1 Aplicações da Prospeção de Dados no Setor da Saúde

O número de investigações baseadas na PD, desde a sua origem, e enquadradas em qualquer setor económico, cresceram exponencialmente. Nos últimos anos, a aplicação destas técnicas ao setor da saúde tem sido alvo de um elevado interesse, pois pode permitir, por exemplo, fazer face ao crescente volume de dados resultantes das investigações desenvolvidas em laboratório (Phillips-Wren, Sharkey e Dy, 2008; Zhou et al., 2010). Deste modo, apresentam-se de seguida algumas aplicações da PD no setor da saúde, na Tabela 2.1, bem como as suas vantagens e desafios.

A partir da Tabela 2.1 verifica-se que existe alguma variedade de investigações no setor da saúde, com propósitos diferentes mas com objetivos semelhantes. De um modo geral, em todas as investigações pretendeu-se explorar os RES com a expetativa de aumentar o

conhecimento existente sobre, por exemplo, as doenças, os pacientes, o diagnóstico e os tratamentos (Mullins et al., 2006; Phillips-Wren, et al., 2008; Zhou et al., 2010; Yeh, Cheng e Chen, 2011).

Tabela 2.1 - Descrição de alguns estudos na área da saúde envolvendo a PD

Objetivo	**Dados**	**Métodos**	**Referência**
Investigar as potenciais vantagens da análise por PD com o objetivo de descobrir novas associações entre os comportamentos de risco e as doenças	E.g.: diagnóstico, microbiológicos, demográficos e tratamentos (*p*=208; *N*=667000)	Descoberta de padrões, análise preditiva e CliniMiner	Mullins et al., 2006
Identificar relações clínicas e demográficas através das variações dos tratamentos do cancro do pulmão que pudessem ser usadas na gestão dos cuidados de saúde	E.g.: demográficos, tratamentos e número de visitas ao médico (*p*=42; *N*=4365)	Árvores de decisão, regressão logística e redes neuronais	Phillips-Wren et al., 2008
Descobrir propriedades que diferenciam os sintomas, e identificar padrões úteis dos pontos de acupuntura e da combinação de ervas utilizadas nas prescrições clínicas da medicina chinesa para apoio da decisão clínica	E.g.: composição das ervas medicinais, sintomas e terapias (*p*= sem informação; *N*=20000)	Máquinas de vetores de suporte, árvores de decisão, redes bayesianas	Zhou et al., 2010
Aplicar métodos de classificação com o objetivo de construir modelos preditivos para o diagnóstico do AVC	E.g.: diagnóstico, exames físicos e exames ao sangue (*p*=29; *N*=493)	Árvores de decisão, classificador bayesiano e redes neuronais	Yeh et al., 2011

Nota: na segunda coluna o número de variáveis (*p*) e a dimensão da amostra (*N*) estão indicados entre parênteses.

Geralmente, este tipo de investigações envolve um elevado número de características, o que dificulta a gestão da aplicação da PD, constituindo um obstáculo ao seu sucesso (Piramuthu, 2004). Por vezes, considera-se um elevado número de variáveis (Mullins et al., 2006) com o intuito de caracterizar o melhor possível as observações em estudo; outras vezes, a

seleção das variáveis é feita por profissionais de saúde com base no seu conhecimento e experiência (Yeh et al., 2011).

O processo de seleção das variáveis de interesse não segue nenhuma regra; é responsabilidade do investigador definir as variáveis de interesse que melhor se adequam ao objetivo da investigação. Como já referido, pode-se recorrer a métodos de PD ou aos conselhos dos especialistas na área para apoiar a sua decisão (Piramuthu, 2004; Kristianson, Ljunggren e Gustafsson, 2009; Hagar et al., 2014). Quando se tem presente um elevado número de variáveis, pode-se reduzi-lo através da transformação das variáveis iniciais em novas que agreguem a informação de mais do que uma variável inicial (Breault, Goodall e Fos, 2002).

Numa outra perspetiva, seria adequado explorar outras tarefas de PD, nomeadamente a regressão (Piramuthu, 2004; Phillips-Wren et al., 2008) para encontrar um conjunto reduzido de variáveis explicativas importantes e excluir as que não apresentassem relevância para a explicação da variável-resposta. Seria igualmente adequado explorar algumas estratégias de redução de dados que permitem obter um número reduzido de novas variáveis a partir de um elevado número, por exemplo, a análise de componentes principais. Este processo admite uma redução pouco significativa da informação com a intenção de simplificar o processamento dos dados (Hand et al., 2001; Everitt e Hothorn, 2011; Han et al., 2011).

Nas investigações apresentadas na Tabela 2.1, de um modo geral, pretendia-se classificar os tratamentos segundo um determinado diagnóstico ou prever o diagnóstico segundo os resultados dos exames realizados. Assim, foram construídos modelos preditivos com o intuito de prever um valor desconhecido de uma variável, do tipo categórica ou quantitativa, segundo os valores conhecidos das restantes variáveis (Hand et al., 2001; Han et al., 2011).

Apesar da complexidade deste tipo de investigação, é interessante verificar o surgimento de novas abordagens de PD enquadradas no setor da saúde, particularmente, a CliniMiner. Esta nova abordagem está relacionada com a associação de acontecimentos com ou sem ocorrência (Mullins et al., 2006). Além de se desenvolverem e explorarem novas abordagens é igualmente relevante a combinação de vários métodos de PD (Mullins et al., 2006; Phillips-Wren et al., 2008; Zhou et al., 2010; Yeh et al., 2011), assim, através desta combinação pretende-se recorrer às vantagens de cada método de modo a alcançar melhores resultados. Pode-se, por exemplo, utilizar árvore de decisão para identificar um subconjunto de variáveis de entrada que seja mais relevante (Phillips-Wren et al., 2008; Yeh et al., 2011), e recorrer a redes neuronais para produzir modelos preditivos com maior precisão, com base nas variáveis selecionadas no modelo anterior (Phillips-Wren et al., 2008).

Uma vez obtidos os resultados e depois de serem automaticamente filtrados, podem ser confirmados manualmente através de referências na literatura médica (Mullins et al., 2006) ou com a participação de médicos experientes no assunto que poderão confirmar a veracidade e utilidade dos resultados para apoio da decisão clínica (Yeh et al., 2011).

2.3 Análise de Dados Espaciais

Principalmente nas últimas duas décadas têm-se verificado progressos evidentes na PD em diversos setores. Devido à sua tecnologia abrangente, pode ser adequada aos diversos tipos de dados existentes, desde que os mesmos tenham a devida importância para o objetivo da investigação. Em particular, foram desenvolvidas metodologias de PD adequadas para processar dados mais complexos, como por exemplo dados espaciotemporais e dados geográficos espaciais, o que constitui uma das principais tendências atuais da PD e também o tema desta seção (Han et al., 2011).

A análise de dados espaciais deriva da PD e foca-se na descoberta de padrões e de conhecimento, a nível geográfico, a partir de métodos adequados aos dados espaciais. Através desta análise pretende-se procurar padrões nos dados que descrevam, por exemplo, as variações das taxas de pobreza em uma zona urbana com base no cálculo das distâncias do centro da região às principais vias rodoviárias. Além da descoberta por padrões no conjunto de dados, pode-se também examinar as relações entre um conjunto de observações espaciais com a finalidade de descobrir quais os subconjuntos dos mesmos que estão espacialmente autocorrelacionados. Também se pode recorrer à análise de dados espaciais para investigar localizações atípicas, i.e., localizações que sejam significativamente diferentes das restantes, através da exploração de áreas locais dentro de uma grande área geográfica (Han et al., 2011).

Relativamente às aplicações da análise de dados espaciais enquadradas no setor da saúde, o maior interesse foca-se no mapeamento da doença com o intuito de investigar a distribuição da doença (Bivand, Pebesma e Gómez-Rubio, 2008). Dentro do mapeamento da doença, existem dois tópicos com maior interesse:

- exibir a variação espacial da incidência de uma doença - este processo fornece uma primeira perspetiva da distribuição espacial da doença que pode auxiliar a detetar áreas em que a doença é particularmente dominante, o que pode levar à deteção de fatores de risco que eram desconhecidos anteriormente;
- localizar a presença de zonas onde o risco tende a ser extraordinariamente mais elevado do que o esperado.

Em relação à aplicação desta análise para o estudo da distribuição geográfica do cancro do pulmão, a Tabela 2.2 apresenta alguns trabalhos desenvolvidos nesta área.

Tabela 2.2 - Descrição de estudos sobre o cancro do pulmão envolvendo a análise de dados espaciais

Objetivo	**Dados**	**Métodos**	***Software***	**Referência**
Investigar a relação entre quatro fontes de poluição ambiental e o risco de cancro do pulmão, nos indivíduos do sexo masculino, em dois períodos temporais	E.g.: idade, probabilidade de exposição a agentes cancerígenos (baseado na tipologia do emprego)	Modelos espaciais baseados na frequência de eventos por unidade de área	ARC/Info 6.1; Gauss 2.2	Biggeri, Barbone, Lagazio, Bovenzi e Stanta, 1996
Descobrir padrões espaciais e localizações dissemelhantes das restantes em relação às taxas de mortalidade do cancro do pulmão	E.g.: código postal, idade, número de casos observados, números de casos esperados	Modelos espaciais neutros e estatística local de Moran (LISA)	ClusterSeer	Goovaerts e Jacquez, 2004
Estudar a distribuição geográfica da incidência do cancro do pulmão em 2011 para fornecer pistas científicas para a prevenção do cancro do pulmão	E.g.: sexo, idade, taxas de incidência, fatores geográficos meteorológicos (taxa de cobertura florestal, índice de qualidade do ar, etc.)	Análise de correlação (Spearman) e regressão	Arcview 3.0; SPSS	Lin et al., 2013
Encontrar a relação entre as mortes por cancro do pulmão, entre 2004 e 2010, com a acumulação de radão nas habitações, na população com idade igual ou superior a 30 anos	E.g.: idade, sexo, ano do falecimento, residência, níveis de radão interior e exterior	Coeficiente de correlação *I* de Moran	ArcGis	Hinojosa de la Garza et al., 2014

De um modo geral, os trabalhos apresentados na Tabela 2.2 pretendem estudar o comportamento do cancro do pulmão a nível geográfico, relacionando a sua mortalidade com fatores ambientais tais como: fontes de poluição (Biggeri et al., 1996), taxas de cobertura florestal, níveis de precipitação anual, índice de qualidade do ar (Lin et al., 2013) e níveis de radão (Hinojosa de la Garza et al., 2014), ou através da análise da distribuição e correlação das taxas de mortalidade a nível geográfico (Goovaerts e Jacquez, 2004).

Com estes estudos podemos verificar, por exemplo, que uma taxa de cobertura florestal baixa, um índice de qualidade do ar fraco e um nível de precipitação anual baixo, estão relacionados com a incidência do cancro do pulmão na região em estudo (Lin et al., 2013).

Além da relevância do estudo dos fatores de risco em cruzamento com as taxas de incidência, por exemplo - uma vez que se tem verificado relação constante entre os dois pontos, é igualmente relevante conhecer a região em estudo. O facto de se saber que a região tem uma elevada concentração de indústrias (Biggeri et al., 1996) ou que se situa numa zona próxima de minas de urânio (Hinojosa de la Garza et al., 2014), pode auxiliar na interpretação dos resultados obtidos. Pode-se chegar à conclusão anterior uma vez que, raramente, neste tipo de situações, existe aleatoriedade de acontecimentos. Desta forma, geralmente tem-se presente um padrão espacial neste tipo de dados devido à interação entre localizações. Para este tipo de cenários existem métodos espaciais neutros que tornam-se relevantes para remover esta interação (Goovaerts e Jacquez, 2004).

Em suma, a análise de dados espaciais pode contribuir para reforçar o conhecimento existente sobre uma determinada doença, a sua distribuição geográfica, e por outro lado fornecer matéria para futuras investigações.

2.4 Análise de Agrupamentos

Neste tópico iremos debruçar-nos sobre a análise de agrupamentos, uma das vertentes da PD inserida nos métodos de aprendizagem não supervisionada, e algumas das suas aplicações a investigações sobre o cancro do pulmão.

O conceito base da análise de agrupamentos é, de certa forma, primitivo e adequado às capacidades de perceção da mente humana.

> An Intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar

> objects encountered in the past to the object at hand (Pinker, 1997, p. 163).

Os seres humanos, por exemplo, podem ser agrupados segundo o seu sexo, masculino e feminino, onde cada grupo tem características específicas comuns - aspetos físicos e comportamentais - que podem identificar facilmente o grupo, e ainda realçar certas propriedades que não seriam visíveis sem este agrupamento (Needham, 1965; Everitt, Landau, Leese e Stahl, 2011).

A análise de agrupamentos, como já referido, enquadra-se na aprendizagem não supervisionada, na qual inicialmente não são impostos objetivos específicos e desconhece-se a classe das observações à partida, estando presente a expectativa de encontrar resultados simultaneamente interessantes e inesperados. Este tipo de análise pretende alcançar um modelo descritivo que apresente de forma conveniente, as características principais dos dados. De um modo geral, o intuito desta análise consiste em formar conjuntos nos quais os elementos pertencentes ao mesmo grupo sejam, o mais possível, semelhantes entre si e diferentes dos outros elementos. Para medir o grau de parecença entre uma observação e outra recorre-se a medidas de semelhança ou distância, dependendo da tipologia das variáveis (Hand et al., 2001; Everitt et al., 2011; Han et al., 2011).

Esta análise proporciona uma correlação com estruturas de dados que não seria possível pelos métodos tradicionais de representação gráfica e estatísticas descritivas. A análise de agrupamentos pode ser utilizada: como ferramenta de pré-processamento de dados para obter um sumário de conjunto, dado que, com frequência, os investigadores se deparam com grande quantidade de informação; para encontrar subconjuntos de dados; e ainda para deteção de valores atípicos uma vez que, geralmente, estes são detetados por se encontrarem muito afastados de qualquer grupo (Hand et al., 2001; Everitt et al., 2011; Han et al., 2011).

Existe alguma variedade de métodos da análise de agrupamentos, onde cada método possui algumas propriedades que o distingue dos restantes. Seguidamente, descreve-se algumas características dos principais métodos da análise referida (Hand et al., 2001; Han et al., 2011).

- **Particionamento -** a fase inicial deste método consiste em organizar as observações a um certo número de grupos definido à partida pelo investigador. Posteriormente, o algoritmo realoca as observações de forma a otimizar a função de qualidade do método escolhida (e.g.: variação intragrupo, variação intergrupos), ou seja, em cada interação existe o mesmo número inicial de

grupos, simplesmente as observações são realocados segundo uma medida de avaliação do particionamento. Dada esta natureza deste método, geralmente, obtém-se um particionamento dos dados mais eficaz, em comparação com outros métodos. Uma vez que o algoritmo escolhe um número definido à partida de centros aleatoriamente selecionados, a partição das observações final pode ser diferente em cada execução do algoritmo, dando origem a resultados diferentes para o mesmo conjunto de dados e método.

- **Hierárquico -** consiste em decomposição hierárquica das observações de um conjunto de dados, segundo um certo critério. Neste tipo de abordagem não é necessário que seja conhecido à partida o número final de grupos. Contudo, uma vez associada a observação a um certo grupo, não existe a possibilidade de alterar a sua posição nas interações seguintes, mesmo que essa ação melhorasse a qualidade do agrupamento. O método hierárquico divide-se em dois métodos segundo a construção da estrutura hierárquica:
    - **método aglomerativo** – este começa por considerar cada observação como sendo um só grupo, depois, em cada interação seguinte agrupa os grupos segundo certas medidas de distância (e.g.: vizinho mais próximo, vizinho mais longe) até que todas as observações formem um único agrupamento ou que outro critério de paragem seja satisfeito;
    - **método divisivo** – este mantém a mesma ideia mais no sentido inverso, ou seja, o algoritmo começa com um único grupo contendo todas as observações, e seguidamente é procurada uma divisão que não prejudique o critério escolhido para avaliar a qualidade do método até que cada grupo seja constituído por uma única observação (ponto de partido do método aglomerativo).
- **Baseados em Densidade -** este tipo de algoritmos baseia-se na densidade das observações numa determinada localização. Pode-se considerar "densidade" o número de observações pertencentes a uma certa vizinhança, formando-se os grupos através da agregação de zonas em que o número de observações na vizinhança exceda algum parâmetro pré definido.
- **Baseados em Grelha -** geralmente estes métodos são usados na análise de dados espaciais, onde a sua ideia base consiste na agregação das observações que estão dentro de uma estrutura de grelha de acordo com um conjunto de

atributos estatísticos provenientes dos cálculos e testes estatísticos efetuados nos dados. Neste caso não se aplicam medidas de distância mas sim segundo a sua densidade e a qualidade do agrupamento oscila consoante o tamanho da grelha.

Uma vez apresentado o conceito geral da análise de agrupamentos será oportuno indicar algumas investigações que envolveram a aplicação de métodos de agrupamento no estudo do cancro do pulmão, como se pode observar na Tabela 2.3.

Tabela 2.3 - Descrição de alguns estudos envolvendo métodos de agrupamento

Objetivo	**Método**	**Referência**
Identificar a localização do maior gene supressor envolvido no cancro do pulmão	Particionamento	Girard, Zöchbauer-Müller, Virmani, Gazdar e Minna, 2000
Análise das expressões de níveis de mRNA correspondentes a 12 600 sequências de transcrição em 186 amostras de cancro do pulmão	Agrupamento hierárquico	Bhattacharjee et al., 2001
Avaliar a prevalência de sintomas nos pacientes com cancro do pulmão para identificar grupos de pacientes agrupados segundo a magnitude de sintomas, e comparar os grupos em relação à qualidade de vida dos pacientes	Agrupamento hierárquico aglomerativo	Franceschini, Jardim, Fernandes, Jamnik e Santoro, 2013
Explorar as diferenças entre indivíduos asiáticos fumadores e não-fumadores com cancro do pulmão entre 2002 e 2006	Agrupamento hierárquico	Krishnan et al., 2014

Na Tabela 2.3 verifica-se uma diversidade de aplicações dos métodos de agrupamento. Através da sua utilização pode-se descobrir, por exemplo, pontos ativos com elevada frequência de perda de *allelic* em cada tipo de cancro no pulmão (Girard et al., 2000) ou concluir que o cruzamento de dados de expressões de perfil com parâmetros clínicos podem ajudar no diagnóstico do cancro do pulmão (Bhattacharjee et al., 2001).

Ao ampliar a investigação de modo a incluir dados demográficos (e.g.: idade e sexo) ou comportamentos de risco (e.g.: hábitos tabágicos e níveis de qualidade de vida) aos dados

clínicos e de ADN temos acesso a outro tipo de resultados e informação: verifica-se que a avaliação de grupos de sintomas serve de ferramenta para medir a qualidade de vida dos pacientes com cancro do pulmão (Franceschini et al., 2013), ou pela análise dos agrupamentos obtida, verificar que uma pequena proporção de fumadores foram classificados no grupo onde predominam os indivíduos que nunca fumaram podendo, desta forma, dar origem a futuras investigações para perceber a razão desta classificação (Krishnan et al., 2014).

Contudo, a inclusão deste tipo de variáveis implica um processo de pré-processamento de dados mais moroso devido à inconsistência (e.g.: alguns pacientes terem o histórico de hábitos tabágicos incompleto) e sensibilidade dos dados, uma vez que as observações referem-se a pacientes e não existe, na prática, uma regra fixa para o procedimento mais correto. Dependendo do objetivo da investigação, fica a cargo do investigador determinar a melhor solução para este tipo de problemas.

Os métodos apresentados na Tabela 2.3, de uma forma geral, são métodos de agrupamento hierárquico dado que não requerem a determinação, à partida, do número de agrupamentos que se pretende. Apesar desta vantagem dos métodos de agrupamento hierárquico, os métodos de particionamento obtêm melhores resultados mas requerem que seja definido, à partida, o número de grupos. Para suportar esta decisão pode-se recorrer a métodos estatísticos e de inferência que procuram o número ótimo de grupos para um dado conjunto de dados. Assim, através desta combinação de métodos pode-se apoiar as decisões tomadas ao longo da investigação e ainda aumentar a segurança dos resultados obtidos (Everitt et al., 2011).

2.5 Definição de Cancro

Neste tópico pretende-se, numa breve descrição, enquadrar a origem e desenvolvimento do cancro, e em particular salientar alguns fatores que podem levar ao seu surgimento. Ainda será apresentado o caso particular do cancro do pulmão, bem como algumas das suas características de modo a facilitar a compreensão desta doença.

2.5.1 Origem e Desenvolvimento do Cancro

Uma alteração no ciclo de vida da célula pode dar origem ao surgimento do cancro. O ciclo de vida da célula é um processo do organismo bastante rigoroso (Denoix, 1977; Del Giglio, 1996; Lodish et al., 2000), o que leva à necessidade de, primeiramente, descrever o normal processo de divisão celular.

Cada célula tem no seu ADN a informação sobre a sua função no organismo, i.e., cada célula deve ainda assegurar os meios para manter a sua existência e transmitir às suas descendentes todas as informações que assegurem a continuação do desempenho do seu papel na sua localização específica, como ilustra a Figura 2.3. O processo de replicação das células está meticulosamente programado, sendo regulado pelas necessidades do organismo. Quando o número de células jovens coincide com o número de células que morreram e que foram substituídas, o processo de renovação celular termina, evitando-se um excesso de células (Denoix, 1977; Del Giglio, 1996; Lodish et al., 2000).

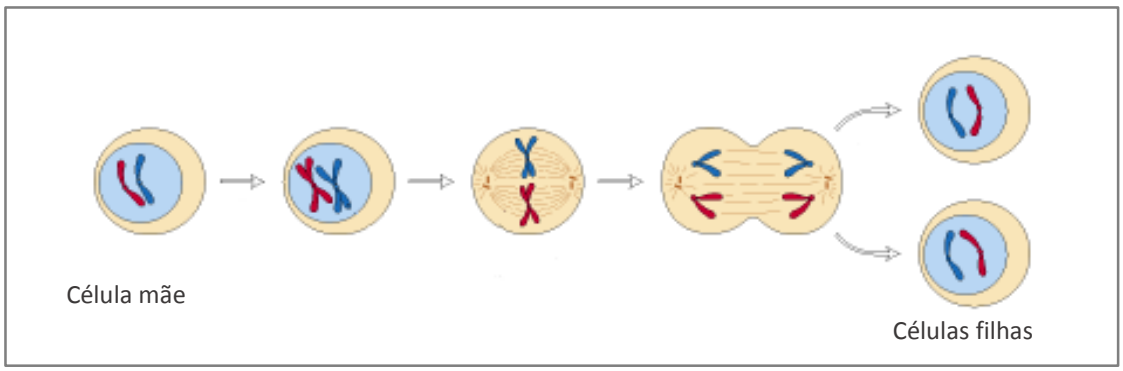

Figura 2.3 - Processo de divisão celular (adptado de Lodisch et al., 2000, p. 11)

O cancro surge quando uma célula deixa de seguir as informações pré-definidas para o seu ciclo de vida e começa a ter um comportamento isolado das células adjacentes, deixando de responder às necessidades do organismo para o qual estava destinada. As características da célula desregulada são transmitidas às células descendentes, que no seu conjunto formam uma massa de células designada "tumor", desencadeando um crescimento e divisão celular desregulados e multiplicando-se a um ritmo mais acelerado do que o habitual. O tumor poderá infiltrar-se nos tecidos adjacentes, danificando a estrutura do órgão onde se encontra e alastrando-se para o exterior do órgão (Denoix, 1977; Del Giglio, 1996; Lodish et al., 2000).

Geralmente, o cancro resulta de mutações acumuladas nas células somáticas, podendo resultar da acumulação de mais ou menos mutações, ao longo dos anos. Assim, podem decorrer vários anos até que ocorram as mutações acumuladas num volume considerável para que o cancro seja possível de diagnosticar, o que leva a que o cancro seja frequentemente considerado

uma doença relacionada com os idosos. Tal como a gravidade do cancro aumenta com a idade, também a gravidade do tumor aumenta com a ocorrência de metástases (Denoix, 1977; Lodish et al., 2000).

Fumar é o comportamento de risco com maior ocorrência nos novos casos de cancro, seguido do consumo de álcool, alimentação não saudável e sedentarismo. Também os raios ultravioleta, radiação ionizante, produtos químicos como o amianto e o arsênico, infeções provocadas por certos vírus, bactérias ou parasitas combinadas com fatores genéticos podem desencadear cancro. Ao nível mundial, mais de 30% das mortes por cancro podem ser atenuadas se os comportamentos e fatores de risco forem alterados ou se o seu diagnóstico for precoce (WHO, 2015).

2.5.2 Cancro do Pulmão

O cancro do pulmão surge quando existe uma desregulação no desenvolvimento espectável de uma célula pertencente ao pulmão, que começa a ter um comportamento individual e é desencadeado um processo de divisão celular à margem das necessidades do organismo (Fundação Portuguesa do Pulmão [FPP], s.d.; CancerCare, 2015a).

Pensa-se que tanto fatores exteriores e independentes do indivíduo, como próprios do seu ADN e ainda comportamentos de risco individuais estejam relacionados com o aparecimento do cancro do pulmão (Figura 2.4) (FPP, s.d.; Ismael et al., 2010; CancerCare, 2015b).

Um estudo epidemiológico do cancro do pulmão em Portugal Continental entre 2000 e 2002, realizado pela Comissão de Trabalho da Pneumologia Oncológica (Parente et al., 2007), verificou os seguintes as seguintes constatações: cerca de 23,7% dos pacientes incluídos na amostra (N = 4396) eram não fumadores; o carcinoma de células escamosas e o CPPC, que são tumores relacionados com o tabaco, apresentaram uma diminuição sucessiva ao longo dos anos, com apenas 23,2% dos pacientes a serem diagnosticados em estádios eventualmente cirúrgicos; por fim, o adenocarcinoma, comparativamente ao carcinoma de células escamosas, tem uma maior associação com o sexo feminino.

Esse estudo permitiu também observar que, de um modo geral, a proporção de fumadores era significativamente superior ao de não fumadores ou ex-fumadores independentemente do tipo de carcinoma. Não se verificaram relações significativas entre o hábito tabágico e o estádio da doença, entre os hábitos de fumo e a região do diagnóstico dos pacientes, nem entre a evolução da doença e a região do diagnóstico dos pacientes.

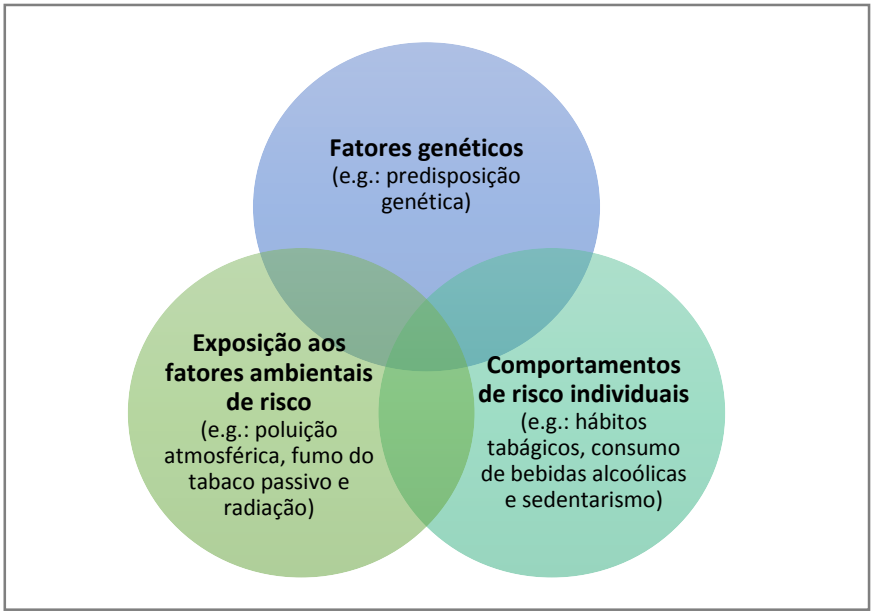

Figura 2.4 - Fatores de risco do cancro do pulmão (FPP, s.d.; Ismael et al., 2010; CancerCare, 2015b)

Estes factos foram agravados com a conclusão de que na maior parte dos pacientes o diagnóstico é tardio, tornando manifestas as dificuldades de um diagnóstico precoce capaz de salvar muitas vidas (Parente et al., 2007).

3 METODOLOGIA

Neste capítulo descreve-se a forma de obtenção, organização, tratamento e processamento dos dados, a saber: o *software* escolhido; pré-processamento de dados; inferência estatística; análise de dados espaciais; e análise de agrupamentos.

3.1 ESCOLHA DO *SOFTWARE*

Como tem sido mencionado ao longo deste documento, este estudo compõe-se essencialmente por duas análises: análise de agrupamentos e análise de dados espaciais. Para a escolha de software utilizou-se os critérios de licença livre, aceitação pela comunidade e polivalência.

Para a análise de agrupamentos, pode-se, por exemplo, utilizar o *software* R, que é dedicado a computação estatística e gráfica, de licença livre, e disponibiliza métodos clássicos e inovadores de análise estatística (The R Foundation, s.d.). Por outro lado, pode-se recorrer a ferramentas específicas para aprendizagem automática como o WEKA. Esta ferramenta agrega, entre outras funcionalidades, uma coleção de algoritmos de aprendizagem automática para tarefas de PD (The University of Waikato, s.d.).

Para a análise de dados espaciais utiliza-se tradicionalmente sistemas de informação geográfica, como se exemplificou na Tabela 2.2. Para além das ferramentas já referidas, um dos programas de licença livre recomendados pelos geógrafos é o QuantumGis (QGIS, s.d.) que permite criar, editar, visualizar e publicar informações geoespaciais. Por outro lado, o GeoDa é igualmente um sistema de informação geográfica, também de licença livre, que disponibiliza, entre outras, funcionalidades de mapeamento para análise exploratória de dados espaciais (GeoDa Center, s.d.). O ambiente interativo constitui a sua principal característica, o que se traduz numa utilização simples e intuitiva por parte do utilizador. Em contraste com o GeoDa, fez-se um esforço para a criação de pacotes no *software* R que permitisse estender o mesmo com métodos de análise de dados espaciais, tornando-o, desta forma, numa ferramenta igualmente adequada para este tipo de análise (Anselin, Syabri e Kho, 2006; Bivand et. al, 2008).

Perante as opções disponíveis, decidiu-se utilizar unicamente o *software* R (versão x64 3.2.2), uma vez que esta ferramenta permite desenvolver todas as tarefas incluídas neste

estudo, é amplamente utilizado tanto pela comunidade de análise estatística como pela comunidade de prospeção de dados, e é de uso livre.

3.2 PRÉ-PROCESSAMENTO DE DADOS

O pré-processamento de dados é uma fase do processo de PD que serve principalmente para organizar e preparar a informação para as etapas seguintes. Estima-se que 80% da duração total deste tipo de análise seja gasto na fase referida, em particular, devido à existência de dados incompletos e incoerentes.

Contudo, pode-se otimizar este tempo ao realizar-se uma boa seleção das variáveis a incluir no estudo, tendo em conta a relevância e redundância de cada variável, bem como o planeamento adequado das estratégias de limpeza de dados (Hand et al., 2001; Piramuthu, 2004; Everitt e Hothorn, 2011).

Neste tópico apresenta-se os dados e as respetivas fontes, os desafios e as estratégias utilizadas na limpeza e construção de dados segundo a sua estrutura e, ainda, a integração de dados realizada.

3.2.1 Extração de Dados

De um modo geral, pretende-se captar as características mais relevantes de um conjunto de observações através de um conjunto de variáveis. Este último conjunto pode não incluir todas as variáveis existentes, mas serve para selecionar as que providenciam uma maior quantidade de informação não redundante e pertinente para o conceito em estudo.

Como o conjunto de variáveis disponível pode não ser o melhor para descrever as características principais das observações, recomenda-se uma análise entre o conjunto de variáveis selecionado e cada uma das variáveis individualmente, interrogando-se do posicionamento de cada variável no conjunto. Uma estratégia para atenuar a complexidade deste processo passa por agrupar as variáveis em grupos de características principais. Considere-se o seguinte exemplo, supõe-se que um dos conjuntos de variáveis se referem às caraterísticas dos pacientes, e que uma das variáveis existentes é a idade do paciente. A relevância desta variável (e, posteriormente, das restantes) deverá ser individualmente avaliada, dentro e fora dos grupos formados (Breault et al., 2002; Piramuthu, 2004; Phillips-Wren et al., 2008; Zhou et al., 2010; Everitt e Hothorn, 2011).

Com base nas estratégias referidas, recorreu-se a seis fontes de dados distintas sendo cada uma utilizada com um propósito específico. Utilizou-se os repositórios: do ROR-Sul, do

Instituto Nacional de Estatística (INE), da PORDATA, da Agência Portuguesa do Ambiente (APA), do Inquérito Nacional de Saúde (INS) e da Direção Geral do Território (DGT), como ilustra a Figura 3.1.

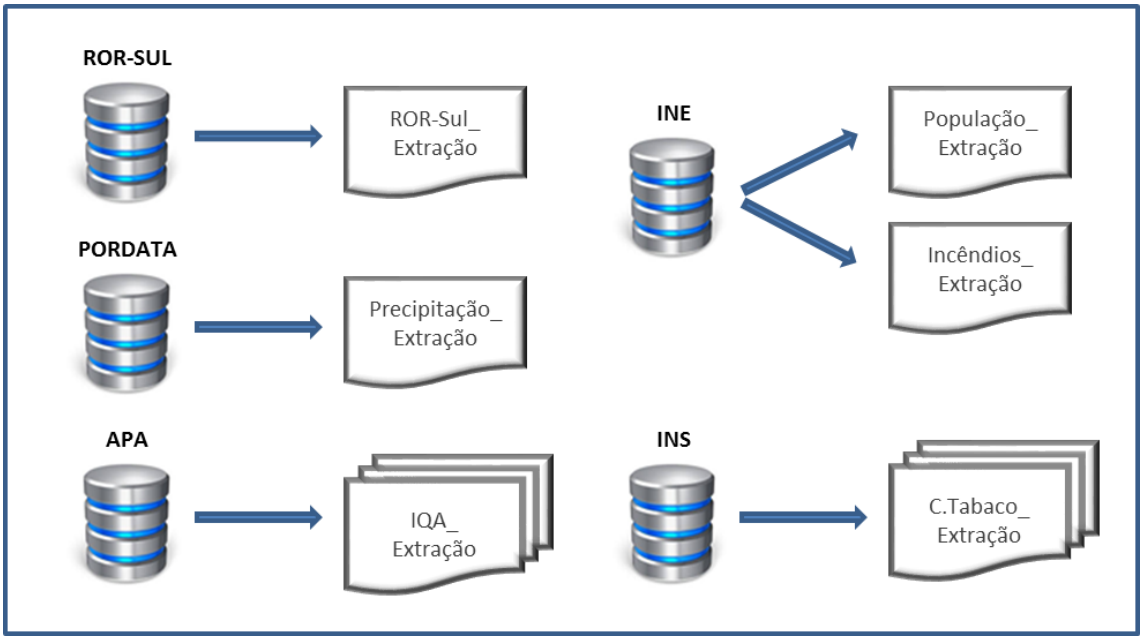

Figura 3.1 - Diagrama do processo de extração de dados

De seguida apresentam-se os dados extraídos de cada fonte mencionada.

Extração de Dados do ROR-Sul

Pretendeu-se extrair um conjunto de variáveis que captassem características relevantes para este estudo - informações sobre os sintomas, o paciente, o diagnóstico, o tumor e os tratamentos realizados. A partir da coleção de variáveis disponíveis (consultar o Anexo 2 para ver a lista completa) que poderiam ser incluídas, e segundo as recomendações dos profissionais do ROR-Sul, organizaram-se três categorias:

- características do paciente -
  - sexo,
  - data de nascimento,
  - distrito de residência no diagnóstico,

    - concelho de residência no diagnóstico,
    - freguesia de residência no diagnóstico,
    - estado vital do paciente no último contato,
    - data de último contacto;
- características do diagnóstico -
    - data de diagnóstico,
    - classificação TMN de tumores malignos clínica,
    - classificação TNM de tumores malignos patológica,
    - estádio[1];
- características do tumor -
    - topografia,
    - morfologia,
    - grau de diferenciação do tumor,
    - estádio na apresentação.

Salienta-se que a base de dados do ROR-Sul está diariamente a ser atualizada e, como tal, ao longo deste estudo realizaram-se duas extrações a partir da mesma, em períodos diferentes, de modo a elevar a qualidade deste estudo.

Extração de Dados do Instituto Nacional de Estatística

Recorreu-se a este repositório com dois propósitos: construir variáveis relevantes para a análise de dados espaciais (nomeadamente as taxas de incidência), e explorar potenciais relações entre os RES e os fatores ambientais. Para tal, através da plataforma *online* do INE extraíram-se dados relativos a:

- estimativas da população residente, por concelho de residência (somente os concelhos abrangidos pelo ROR-Sul), sexo e faixa etária, nos anos de 2012 e 2013 (INE, 2015a);
- número de incêndios florestais, por concelho pertencente à região ROR-Sul, entre 2000 e 2013 que corresponde ao período de anos disponível (INE, 2015b).

[1] A forma estadio é também passível de ser utilizada.

Extração de Dados da PORDATA

Com o propósito de completar a informação sobre os fatores ambientais, extraíram-se, da plataforma *online* da PORDATA, dados da precipitação anual (mm) desde 1960 até 2013, das quatro estações meteorológicas abrangidas pela região ROR-Sul: Lisboa, Beja, Faro e Funchal (PORDATA, 2015).

Extração de Dados da Agência Portuguesa do Ambiente

O portal da APA disponibiliza uma base de dados *online* com informações sobre a qualidade do ar a nível nacional. O Índice de Qualidade do Ar (IQA) permite uma classificação do estado de qualidade do ar de uma determinada zona do país, sendo cada zona classificada numa das categorias pré-definidas, todos os dias do ano. Desta plataforma extraíram-se dados relativos à qualidade do ar na região ROR-Sul: o histórico anual que consiste no número de dias em cada uma das categorias, entre 2004 e 2013 (APA, s.d.).

Extração de Dados do Inquérito Nacional de Saúde

Dado que o conjunto de variáveis extraídas da base de dados do ROR-Sul, que caracterizam os pacientes com cancro do pulmão, excluem dados relativos aos comportamentos de risco de cada paciente, recorreu-se ao último INS 2005/2006 com o objetivo de tentar completar esta informação. Assim, extraiu-se a proporção da população residente em Portugal em 2005/2006, com 15 ou mais anos, por sexo e grupo etário, por região (LVT, Alentejo, Algarve e RAM) e por consumo de tabaco. Neste, considera-se "nunca fumadores" e "fumadores" (que inclui ex-fumadores e fumadores atuais, ocasionalmente ou diariamente) (Instituto Nacional de Saúde Dr. Ricardo Jorge, 2010).

Extração de Dados da Direção Geral do Território

A análise de dados espaciais pressupõe a integração de dados não espaciais, relevantes para o estudo, com dados espaciais, i.e., dados geográficos. Deste modo, a partir da plataforma *online* da DGT, fez-se a extração dos ficheiros com os troços e limites administrativos dos concelhos de Portugal, correspondentes ao Continente e à Região Autónoma da Madeira. Esta informação está de acordo com as reorganizações mais recentes da Carta Administrativa Oficial de Portugal (DGT, 2015).

3.2.2 Limpeza e Construção de Dados

No processo de PD, é igualmente importante uma sensata extração de dados e a sua preparação.

Numa primeira fase pode-se apontar a presença de valores omissos no conjunto de dados como um dos grandes desafios da PD, uma vez que alguns dos seus métodos são sensíveis a este tipo de características. Nesta situação, a melhor estratégia está dependente da sensibilidade do conceito em estudo. Além deste desafio, torna-se fundamental validá-los como consistentes, isto é, se estão tecnicamente corretos para serem usados na análise pretendida.

Deste modo, existem duas preocupações essenciais: a primeira consiste em validar os valores admissíveis de cada variável, verificando se não apresenta valores impossíveis de acordo com o seu significado (da variável); a segunda prende-se com informações contraditórias num conjunto de variáveis para a mesma observação (Cody, Ed, Wood, Medical, s.d.; de Jonge e van der Loo, 2013).

O conjunto de dados em estudo pode ser dividido em três categorias: RES extraídos do repositório do ROR-Sul; dados de fatores ambientais da região ROR-Sul e comportamentos de risco dos indivíduos; e dados de incidência.

Registos Eletrónicos de Saúde

Por via da análise exploratória dos dados extraídos do ROR-Sul, verificou-se que no conjunto de dados existiam 1033 valores omissos (3%), que 127 dos pacientes (13%) tinham características com valores omissos, e existia uma percentagem significativa de valores omissos em algumas variáveis (e.g.: a variável respeitante ao estádio do tumor no momento do diagnóstico tinha 372 valores omissos - 39%).

Dado que neste estudo não se pretendia remover observações nem preencher os dados omissos através de técnicas de simulação e estimação (Kristianson, Ljunggren e Gustafsson, 2009), optou-se por considerar apenas as variáveis que não apresentassem valores omissos (Everitt e Hothorn, 2011).

Na Tabela 3.1 apresentam-se as principais características das variáveis em estudo, relativas ao paciente e ao diagnóstico. Confirmou-se se os valores observados das variáveis eram valores plausíveis para a variável em questão. Validou-se igualmente a conformidade entre a data de nascimento e a data de diagnóstico, e entre o distrito de residência e o concelho de residência no momento do diagnóstico. Neste último caso utilizou-se como referência a

informação territorial extraída da DGT (DGT, 2015). Uma vez que não se encontrou nenhum caso inválido não houve necessidade de correção, caso contrário o mesmo teria sido corrigido.

Tabela 3.1 - Características do paciente e do diagnóstico

Nome da Variável	**Descrição**	**Tipo de Variável**	**Valores Admissíveis**
Sexo	Sexo	Texto	"Feminino" e "Masculino"
DtNasc	Data de nascimento	Data (dd-mm-aaaa)	Qualquer data válida
DistritoDiag	Distrito de residência no diagnóstico	Texto	Qualquer distrito pertencente à região ROR-Sul
ConcelhoDiag	Concelho de residência no diagnóstico	Texto	Qualquer concelho pertencente ao distrito de diagnóstico
DtDiag	Data de diagnóstico	Data (dd-mm-aaaa)	Qualquer data válida entre 01/01/2013 e 30/06/2013
Estado	Estado de vida	Texto	"Falecido" e "Vivo"

As características do tumor estão codificadas na base de dados do ROR-Sul segundo o sistema de classificação das doenças oncológicas definido pela *Internation Classification of Diseases for Oncology – Third Edition*. Esta classificação define de forma precisa a localização anatómica, os tipos histológicos dos tumores e ainda o comportamento dos mesmos (Fritz et al., 2000; ROR-Sul, 2014).

O código topográfico identifica o local de origem do tumor, i.e., agrega a informação sobre a sua localização principal (e.g.: pulmão) e a sublocalização (e.g.: brônquio principal).

O código morfológico é composto por um conjunto que três códigos referentes à identificação do código morfológico, ao tipo de célula do tumor e ao comportamento celular do tumor. O grau de diferenciação dos tumores malignos e sólidos descreve o quanto o tumor se assemelha ao tecido normal no qual teve origem (Fritz et al., 2000).

Relativamente à classificação do estádio do tumor, o *AJCC Cancer Staging Manual* é o guia de referência seguido pelo ROR-Sul (ROR-Sul, 2014). Esta classificação no momento do diagnóstico torna-se fundamental para a definição dos tratamentos mais apropriados, a partir do histórico de pacientes anteriores no mesmo estádio. Este refere-se, principalmente, à componente de presença ou ausência de metástases (Edge et al., 2010).

Na Tabela 3.2 pode-se observar os valores admissíveis para cada variável mencionada anteriormente. Tal como para as variáveis anteriores, obteve-se as frequências das variáveis seguintes de modo a validar os valores observados.

Tabela 3.2 - Características do tumor

Nome da Variável	**Descrição**	**Tipo de Variável**	**Valores Admissíveis**
Top	Topografia	Código alfanumérico	"C34.0", "C34.1", "C34.2", "C34.3", "C34.8", "C34.9"
TopDesc	Designação da codificação da topografia	Texto	"Brônquio principal", "Lobo superior do pulmão", "Lobo médio do pulmão", "Lobo inferior do pulmão", "Múltiplas subcategorias do pulmão", "Pulmão SOE[2]"
Morf	Morfologia	Código alfanumérico	M 8000-9989 / 3
MorfDesc	Designação da codificação da morfologia	Texto	E.g.: "Adenocarcinoma com subtipos mistos"
GrauDif	Grau de diferenciação do tumor	Texto	"Bem diferenciado", "Moderadamente diferenciado", "Pouco diferenciado", "Indiferenciado", "Desconhecido", "Não aplicável"
EstadioApr	Estádio na apresentação do tumor	Texto	"Doença local ou loco-regional", "Doença metastática", "Desconhecido", "Não aplicável"

A partir das variáveis extraídas da base de dados ROR-Sul, e dado que algumas características detinham inúmeras categorias, construíram-se e transformaram-se algumas variáveis por se considerar útil para uma melhor compreensão da estrutura dos dados em estudo.

[2] Sem Outra Especificação

Em relação às características dos pacientes criaram-se as seguintes variáveis: a idade do paciente no momento do diagnóstico, tendo em consideração a data de nascimento e a data de diagnóstico; as faixas etárias que agregam as idades dos pacientes em intervalos de cinco anos exceto no último intervalo; os grupos etários que agregam as idades dos pacientes em intervalos maiores em relação às faixas etárias; e as regiões que são agregações dos distritos da região ROR-Sul mencionados no tópico 1.4 (ROR-Sul, 2014).

Relativamente às características do tumor, criou-se uma variável para agregar tipos histológicos de forma a facilitar a visualização da sua incidência. Esta variável identifica um pequeno número de grupos considerados histologicamente diferentes entre si relativamente à definição de tumores múltiplos (Fritz et al., 2000).

A Tabela 3.3 apresenta as novas variáveis e um resumo das suas características.

Tabela 3.3 - Lista das variáveis construídas no estudo referentes ao paciente e ao tumor

Nome da Variável	**Descrição**	**Tipo de Variável**	**Variáveis de Origem**	**Valores Válidos**
IdadeDiag	Idade do paciente	Inteiro positivo	*DtNas*, *NtDiag*	22 a 91 (anos)
FaixaEtaria	Faixa etária	Texto	*IdadeDiag*	E.g.: “[20,25[”, ”[25,30[” e “[35,40[”
GrupoEtario	Grupos etários	Texto	*IdadeDiag*	“[15,45[”, “[45,55[”, “[55,65[”, “[65,75[”, “[75,100[”
RegiaoDiag	Região de residência	Texto	*DistritoDiag*	“Alentejo”, “Algarve”, “Lisboa e Vale do Tejo”, “Região Autónoma da Madeira”
TTgrupo	Grupos histológicos	Texto	*Morf*	E.g.: "Carcinomas escamosos", "Adenocarcinomas", "Outros carcinomas específicos" e "Carcinomas não específicos (SOE)"

Dados de Fatores Ambientais e Comportamentos de Risco

Em relação aos dados externos, nomeadamente os fatores ambientais, agregou-se a informação em cada variável recolhida, de modo a enquadrar a sua utilização neste estudo uma vez que não se trata de uma análise temporal. É de salientar a existência de uma elevada

proporção de valores omissos nos dados relacionados com os fatores ambientais ao longo dos anos – exemplo: a soma dos dias de uma das classificações do IQA, para uma determinada zona e num determinado ano, não correspondia ao número total de dias do ano, sendo o registo dos restantes dias inexistente.

Ao número de incêndios florestais anuais no período de 2000 a 2013, por concelho, e à precipitação anual entre 1960 e 2013, por região, agregou-se a informação presente em cada ano através da média aritmética das observações anuais para cada concelho e região, respetivamente. Salienta-se, que se considerou, neste último conjunto de dados, que a abrangência territorial de cada estação meteorológica, referida no tópico anterior, seria expandida para a região correspondente - exemplo: a abrangência geográfica da estação meteorológica de Faro seria a região do Algarve. No que se respeita ao histórico anual do IQA optou-se, em primeiro lugar, por obter uma média aritmética que retivesse a informação disponível de todos os anos para cada zona. Posteriormente, construiu-se a variável final considerando duas alternativas:

- Média ponderada para cada zona z da seguinte forma:

  $$IQA_z = Dmt_z \times 0{,}1 + Db_z \times 0{,}15 + Dmd_z \times 0{,}2 + Df_z \times 0{,}25 + Dm_z \times 0{,}3,$$

  onde $Dmt, Db, Dmd, Df\ e\ Dm$ representam o número de dias (anual) em cada uma das seguintes categorias: muito bom, bom, médio, fraco e mau. Esta média ponderada é um índice que pretende agravar o valor de IQA_z à medida que a qualidade do ar decresce numa determinada zona z, i.e., quanto maior o valor de IQA_z maior será o número de dias nas categorias médio, fraco e mau e, consequentemente, pior será a qualidade do ar.
- Duas variáveis que aglomeram as classificações do IQA, ou seja, uma variável, *IQArMelhor*, que é a soma do número médio de dias que foram classificados com muito bom e bom, e outra variável, *IQArPior*, que agrega o número médio de dias das classificações médio, fraco e mau.

Na Tabela 3.4 apresenta-se as variáveis descritas anteriormente relativas aos fatores ambientais.

Tabela 3.4 - Lista das variáveis construídas no estudo relativas aos fatores ambientais

Nome da Variável	**Descrição**	**Tipo de Variável**	**Variáveis de Origem**
MedIncendios	Média do número de incêndios anual	Número positivo	E.g.: *NumIncendios2000*
MedPrecipitacao	Média da precipitação anual	Número positivo	E.g.: *Precipitacao1960*
IQAr	Índice que corresponde à média ponderada do IQA anual	Número positivo	E.g.: *MedMuitoBom*
IQArMelhor	Número de dias, por ano, em que classificou-se as zonas com muito bom e bom	Número positivo	*MedMuitoBom*; *MedBom*
IQArPior	Número de dias, por ano, em que classificou-se as zonas com médio, fraco ou mau	Número positivo	*MedMedio*, *MedFraco*, *MedMau*

Nesta tarefa, de limpeza e construção de dados, não se referiu os dados relativos aos comportamentos de risco dos indivíduos, nomeadamente consumo de tabaco, uma vez que os mesmos não foram usados nas tarefas de PD. Optou-se por não se considerar estes dados porque a sua integração com os RES seria incoerente, na medida que o conjunto de dados constituído pelos RES contém variáveis que descrevem as características de cada paciente. Em contrapartida, os dados de consumo de tabaco referem-se à proporção da população em estudo com determinado tipo de consumo de tabaco.

Dados de Incidência

Neste estudo pretende-se analisar geograficamente a incidência do cancro do pulmão na região ROR-Sul, através da análise de dados espaciais, sendo calculadas para o efeito, as taxas de incidência por concelho, distrito e região, tendo em conta o total da população em cada localização.

A vantagem da utilização das taxas em relação às frequências absolutas centra-se numa aproximação mais real do indicador de incidência, uma vez que regiões mais populosas tendem a ter um maior número de casos, podendo a utilização do número de casos em cada região conduzir a interpretações equivocas (Getis e Ord, 1992).

Antes da construção destas variáveis validaram-se as designações dos concelhos dos dois conjuntos de dados usados nesta tarefa, com recurso aos dados territoriais da DGT. Desta forma, verificaram-se e retificaram-se as designações referentes ao mesmo concelho que não se encontravam no mesmo formato (e.g.: "Calheta" e "Calheta (R.A.M.)"), nos dois conjuntos de dados em causa.

A partir das estimativas da população por concelho obteve-se as estimativas da mesma por distrito e região. A estimativa da população residente a 30 de Junho foi obtida através da média da população residente no final dos anos de 2012 e 2013. As taxas de incidência foram formuladas do modo seguinte (ROR-Sul, 2014):

- taxa de incidência bruta - é o quociente entre o número de novos casos para a localização geográfica l do sexo s e a população residente a 30 de Junho de 2013 na localização geográfica l e do sexo s, por 100 000 habitantes , i.e.,

$$TIB_{ls} = \frac{NC_{ls}}{POP_{ls}} \times 100000;$$

- taxa de incidência específica - é o quociente entre o número de novos casos para a localização geográfica l, do sexo s e do grupo etário (ou faixa etária) i**,** e a população residente a 30 de Junho de 2013 na localização geográfica l, do sexo s e do grupo etário i, por 100 000 habitantes, i.e.,

$$TIE_{lsi} = \frac{NC_{lsi}}{POP_{lsi}} \times 100000.$$

Calcularam-se, ainda, dois tipos de taxas de incidência específicas padronizadas para a idade. Em cada um dos casos considerou-se como população residente o número de efetivos da população padrão europeia e mundial, respetivamente, ambas definidas pelas normas europeias. Estas taxas padronizadas possibilitam fazer-se comparações entre regiões, mas não transmitem o risco real das populações, uma vez que a população padrão utilizada não é real (ROR-Sul, 2014).

A Figura 3.2 apresenta, de uma forma abreviada, os processos apresentados neste tópico de limpeza e organização de dados.

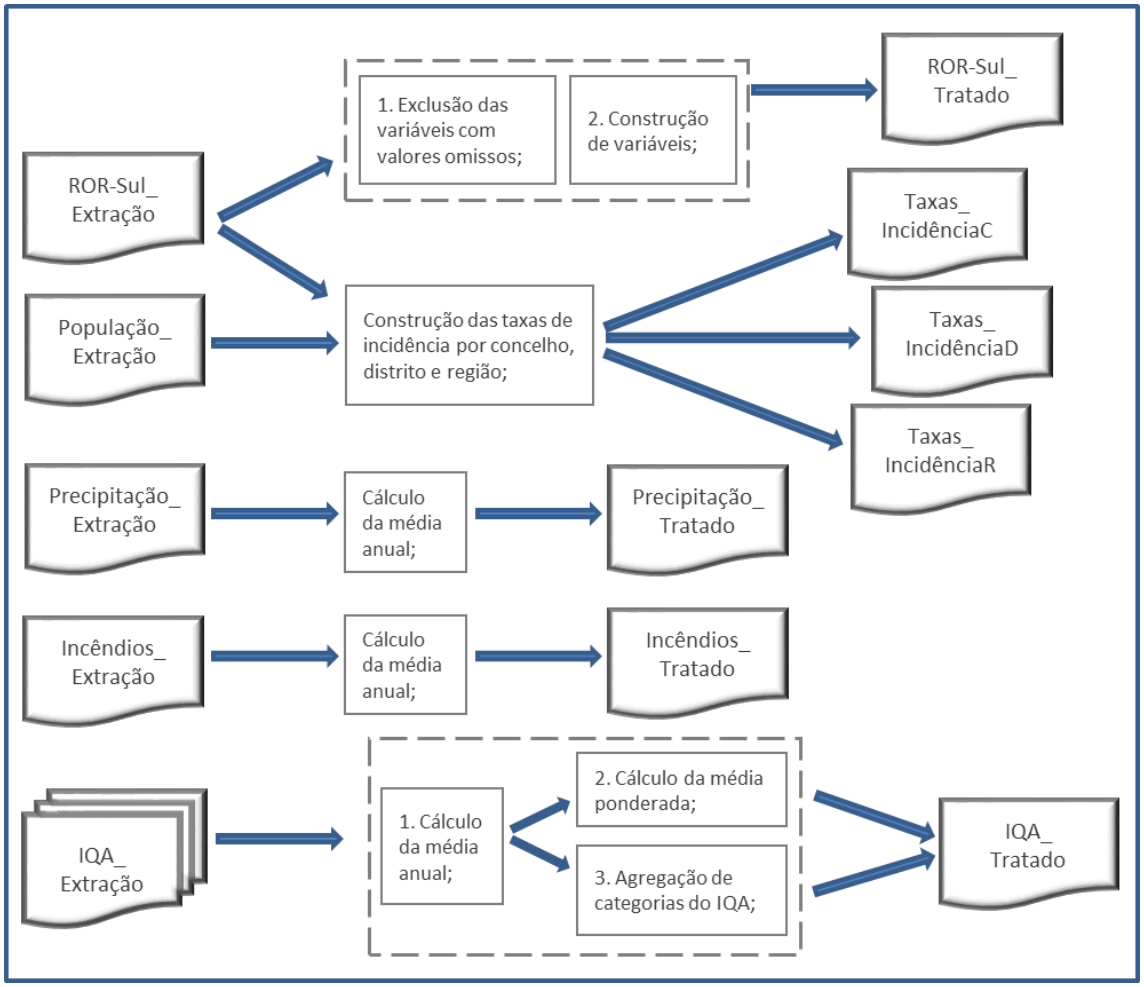

Figura 3.2 - Diagrama do processo de limpeza e organização de dados

3.2.3 Integração de Dados

No processo de integração de dados deve-se ter em atenção que, embora referindo-se às mesmas observações, as variáveis são extraídas de fontes independentes existindo o cuidado de validar se a palavra-chave, que identifica univocamente a observação de cada conjunto de dados, se encontra no mesmo formato. Por vezes a palavra-chave pode ser constituída por um só atributo do conjunto de dados ou pode ser um conjunto dos mesmos, mas a sua utilização é bastante útil e eficaz, transmitindo segurança ao investigador relativamente a erros de integração (Rahm, 2000).

Neste estudo realizaram-se, essencialmente, duas integrações: os dados de fatores ambientais com o conjunto de dados *ROR-Sul_Tratado,* e com o conjunto de dados

Taxas_incidênciaC (Figura 3.2). As informações territoriais foram aplicadas a três níveis diferentes (concelho, distrito e região) como via de integração dos dados em estudo.

Na primeira integração dado que o conjunto de dados *ROR-Sul_Tratado* incluía os três níveis de informação territorial, fez-se a integração da informação através de uma função que compara os pares de conjuntos de dados num dos níveis referidos, como demonstra a Figura 3.3.

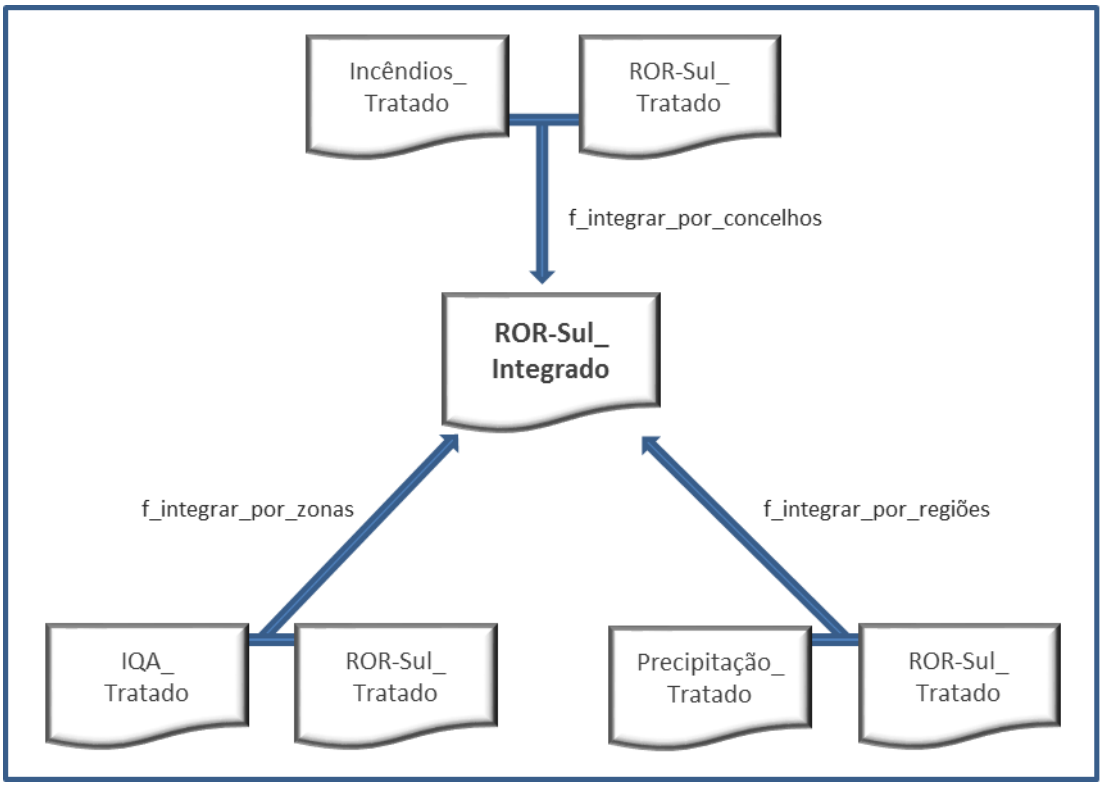

Figura 3.3 - Diagrama do processo de integração dos dados relativos aos pacientes

Como já referido no tópico anterior, verificou-se se as designações dos concelhos da região ROR-Sul se encontravam no mesmo formato em cada conjunto de dados utilizado, nesta tarefa, e retificou-se os casos encontrados. No entanto, no caso particular do conjunto de dados *IQA_Tratado* (Figura 3.3), como as fronteiras territoriais (zona) eram diferentes das fronteiras dos três níveis indicados, agregaram-se localizações de modo a encontrar uma correspondência com as zonas deste conjunto de dados, como por exemplo, a zona "Alentejo Litoral"

corresponde à agregação dos concelhos de Alcácer do Sal, Sines, Grândola, Odemira e Santiago do Cacém.

A segunda integração de dados é análoga à primeira mas, uma vez que o conjunto de dados *Taxas_IncidênciaC* (Figura 3.2) dispunha unicamente da variável territorial *Concelhos*, foi necessário completar este conjunto de dados com a informação territorial presente no conjunto de dados *Info_Território_Extração* (Figura 3.1). Relembra-se que este último conjunto agrega algumas informações territoriais das localizações, nomeadamente as designações dos concelhos e distritos. Desta forma, a partir da informação do distrito correspondente a cada concelho, é possível obter-se a região a que este corresponde, informação necessária para a integração em causa.

Embora tenham sido utilizadas as designações dos concelhos, distritos e regiões (da região de Portugal abrangida pelo ROR-Sul), no processo de integração dos dados em estudo, por existir apenas uma designação para diferentes localizações geográficas, a nível nacional este cenário já não é válido.

3.3 INFERÊNCIA ESTATÍSTICA

A inferência estatística permite deduzir propriedades para a população a partir de propriedades verificadas na amostra. A sua aplicação torna-se relevante quando se pretende estimar, por exemplo, o valor desconhecido de um parâmetro de uma população, ou testar hipóteses relacionadas com a natureza da distribuição da população ou do valor dos parâmetros.

Em particular, os testes de hipóteses consistem, principalmente, no seguinte processo: construir a hipótese que em geral se pretende provar (hipótese alternativa - H_A) e a respetiva hipótese contrária (hipótese nula - H_0); definir a estatística de teste - função da amostra aleatória, cujo valor observado será a base da tomada de decisão (rejeitar ou não H_0); e estabelecer o método de decisão, por exemplo, através do nível de significância (α) e do valor-p (p). O primeiro, α, representa a probabilidade máxima de tomar a decisão de rejeitar uma hipótese nula verdadeira. Este deve ser um valor pequeno e, deste modo, usa-se em geral 0,01; 0,05 e 0,10. O segundo, p, é o valor mais pequeno para o qual a hipótese nula é rejeitada e é obtido através do valor observado da estatística de teste. Assim, rejeita-se H_0 quando o valor de p é menor ou igual a α (Pestana e Velosa, 2008).

Esta tarefa foi incluída neste estudo para melhorar a compreensão das características da população em estudo mas também para testar hipóteses que serão mencionadas de seguida.

Perante a população em estudo surgiram hipóteses em três vertentes: comparação de valores médios; independência em tabelas de contingência; e regressão; que podem ser validadas a partir dos testes apresentados de seguida (consultar o apêndice A para ver o código R respetivo).

Comparação de Valores Médios de Populações Independentes

Pretende-se testar a igualdade do valor médio das idades dos pacientes do sexo masculino e do sexo feminino, ou seja, as seguintes hipóteses:

H_0: a diferença dos dois valores médios é zero

vs.

H_A: a diferença dos dois valores médios é diferente de zero,

através do teste paramétrico *T-Student*, quando as populações, de onde são extraídos os dados, seguem uma distribuição Normal (Murteira e Antunes, 2012).

Para validar a normalidade, aplicou-se inicialmente o gráfico Quantil-Quantil (gráfico Q--Q) - que compara graficamente os quantis empíricos com os quantis correspondentes de uma distribuição teórica (Everitt e Hothorn, 2011; Han et al., 2011). Posteriormente, realizou-se um teste estatístico para aferir se os dados provinham de uma população com distribuição Normal. Os testes de *Lilliefors*, *Shapiro-Wilk* e *Kolmogorov-Smirnov* são talvez os mais usuais, embora o último não seja específico para a distribuição Normal (Razali e Wah, 2011*)*. Entre os dois primeiros testes referidos, optou-se pelo teste de *Shapiro-Wilk* por ser o mais potente (maior capacidade de identificar situações em que a distribuição da população subjacente aos dados não é Normal). O teste estatístico não paramétrico de *Shapiro-Wilk* permite testar as seguintes hipóteses:

H_0: a população segue uma distribuição Normal

vs.

H_A: a população não segue uma distribuição Normal.

Em alternativa ao teste *T-Student*, na eventualidade da hipótese de normalidade ser rejeitada, considerou-se o teste *Kolmogorov-Smirnov* (Pestana e Velosa, 2008; Razali e Wah, 2011). Este teste permite averiguar se a idade dos pacientes do sexo masculino segue uma distribuição semelhante à da idade dos pacientes do sexo feminino e, consequentemente, avaliar as diferenças entre os dois sexos (populações independentes). Desta forma, testa as seguintes hipóteses:

H_0: as duas populações têm a mesma distribuição

vs.

H_A: as duas populações não têm a mesma distribuição.

Independência em Tabelas de Contingência

Uma tabela de contagens, de dupla entrada, diz-se uma tabela de contingência. A partir desta apresentação dos dados pode-se inferir a existência, ou não, de associação entre duas características, geralmente, qualitativas. A independência entre pares de caraterísticas pode ser avaliada aplicando o teste de independência do *Qui-Quadrado*. A estatística de teste, correspondente a este teste, compara as frequências observadas com as esperadas sob a validade de H_0.

O teste de independência do *Qui-Quadrado* pode ser aplicado quando a amostra é suficientemente grande e desde que, sob a hipótese de independência, não existam frequências esperadas inferiores a um e se pelo menos 80% das frequências esperadas sejam superiores a cinco. Nas situações em que o teste do *Qui-Quadrado* não é adequado, devido às suas limitações, a melhor opção será o teste exato de *Fisher* - válido para amostras de qualquer dimensão, embora seja usualmente aplicado no caso de amostras pequenas (Pestana e Velosa, 2008; Murteira e Antunes, 2012).

Os dois testes de independência para tabelas de contingência, referidos no parágrafo anterior, consideram uma estatística de teste com distribuição assintótica de Qui-Quadrado sob H_0, que permite avaliar as seguintes hipóteses:

H_0: As duas variáveis são independentes

vs.

H_A: As duas variáveis não são independentes.

Regressão

Ao nível dos concelhos da região ROR-Sul, pretendeu-se estudar uma possível relação entre uma dada variável em estudo (variável-resposta), como por exemplo, a "taxa de incidência") e as restantes variáveis de interesse, através da análise de regressão múltipla e estimação dos coeficientes associados a cada variável-regressora (Draper e Smith, 1998; Faraway, 2004).

Para avaliar a importância de cada variável-regressora *r*, do modelo de regressão considerado, testam-se as seguintes hipóteses com base numa estatística de teste que, sob H_o, segue uma distribuição *T-Student*:

$$H_0: \beta_r = 0$$

vs.

$$H_A: \beta_r \neq 0.$$

Além do teste estatístico referido considerou-se, como medida de qualidade do ajustamento do modelo, o coeficiente de determinação, R^2. O valor de R^2 encontra-se entre 0 e 1, e indica a percentagem da variável-resposta que é explicada pelo modelo que a relaciona com as variáveis-regressoras. Geralmente, considera-se que o modelo encontra-se bem ajustado quando $R^2 > 0{,}7$; no entanto, este valor de referência é variável e depende do objetivo do estudo (Pestana e Velosa, 2008).

Ainda, como medida de comparação de modelos estimados, considerou-se a medida AIC (*Akaike Information Criteria*) que engloba a precisão e a complexidade do modelo. De um modo geral dá-se preferência ao modelo que apresentar um menor valor de AIC, sendo, à partida, o melhor modelo encontrado entre os estimados (Draper e Smith, 1998).

A aplicação do modelo de regressão tem os seguintes pressupostos: os erros aleatórios são independentes e identicamente distribuídos; os mesmos seguem uma distribuição Normal com valor médio zero e desvio padrão σ (constante); os valores observados dos erros, "resíduos", correspondem à diferença entre o valor observado da variável resposta e o valor estimado pelo modelo; e o modelo só é válido se os resíduos refletirem estas propriedades, ou seja, se assumirem valores muito pequenos, com média zero e distribuídos sem padrão em torno do zero no gráfico "Valores Estimados da Variável Resposta vs. Resíduos" (Draper e Smith, 1998; Pestana e Velosa, 2008).

No seguimento da análise de regressão descrita anteriormente, recorreu-se ao modelo logístico com o intuito de compreender a relação do estádio na apresentação com a idade dos pacientes e os fatores ambientais (Park, 2009). Neste caso consideraram-se apenas duas classes das quatro que compunham esta variável e, portanto, foi transformada numa variável binária.

3.4 Análise de Dados Espaciais

A análise de dados espaciais, como referido no tópico 2.3, permite descobrir padrões espaciais, ou seja, padrões ao longo de uma determinada zona geográfica.

No âmbito da análise geográfica da incidência do cancro do pulmão na região ROR-Sul, decidiu-se aplicar métodos de autocorrelação e associação espacial, inseridos na análise de dados espaciais, para este fim (consultar o Apêndice B para ver o código R respetivo).

Método de Autocorrelação Espacial

A estatística I de *Moran* é uma extensão do coeficiente de *Pearson* e mede a existência de algum padrão ao longo de uma dada área geográfica (Anselin, 1995; Goovaerts e Jacquez, 2004; Bivand et al., 2008). O valor de I resultante de uma amostra de tamanho n é dado pela seguinte expressão:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} {z_i}^2},$$

onde z_i e z_j representam os desvios das observações à média. Define-se $S_0 = \sum_i^n \sum_j^n w_{ij}$ e w_{ij} é o peso espacial entre a observação i e a observação j de uma variável. Considerou-se $w_{ij} = \frac{1}{m}$, onde m representa o número de vizinhos da localização i, i.e., cada vizinho de uma dada localização terá uma ponderação igual às restantes localizações vizinhas, uma vez que não existem evidências para considerar o contrário. Se a observação j não é vizinha de i então $w_{ij} = 0$ e assume-se $w_{ii} = 0$. Portanto, dado que os pesos são uma ponderação padronizada para cada localização iremos obter neste caso $S_0 = n$ (Getis e Ord, 1992; Anselin, 1995). Através desta estatística pode-se testar as seguintes hipóteses:

H_0: não existe autocorrelação espacial entre as observações

vs.

H_A: existe autocorrelação espacial entre as observações.

Uma vez que se trata de um teste não paramétrico e, portanto, com menor potência (menos capacidade para rejeitar H_0 quando esta hipótese é falsa), é recomendado complementar o teste anterior através de um teste de permutações, de modo a providenciar segurança contra erros de inferência.

Método de Associação Espacial

Nos casos em que não se rejeita a hipótese nula referida anteriormente, ou seja, ausência de um padrão espacial global, pode-se estar na presença de associação local, uma vez que a medida I de *Moran* ignora a possibilidade de instabilidade geográfica que pode ser verificada ao nível local. Para colmatar esta situação pode-se recorrer ao indicador local de associação espacial permitindo, desta forma, a decomposição do indicador global de modo a

transparecer a contribuição de cada observação (Anselin, 1995). A sua fórmula é dada pela seguinte expressão:

$$I_i = z_i \sum_{j=1, j\neq i}^{n} w_{ij} z_j,$$

com as suposições análogas ao I de *Moran*, este indicador permite testar as seguintes hipóteses:

H_0: Não existe associação espacial local

vs.

H_A: Existe associação espacial local.

Os valores de I_i positivos e afastados do zero indicam a existência de um agrupamento espacial de valores semelhantes em torno dessa localização e valores negativos afastados do zero indicam a existência de um agrupamento espacial de valores dissemelhantes (Anselin, 1995). Deste modo, este indicador é uma ferramenta versátil que permite:

- identificar agrupamentos locais - identifica padrões locais constituídos por localizações para as quais o valor deste indicador é significativamente semelhante (e.g.: localização com valor elevado da variável de interesse rodeado por localizações com valores igualmente elevados da variável de interesse);
- identificar instabilidade local - identifica padrões locais, nomeadamente, valores muito diferentes da média indicam as localizações que contribuem mais do que a sua parte para a estatística global, podendo desta forma representar zonas atípicas (e.g.: localização com um valor baixo da variável de interesse rodeado por localizações com valores elevados da variável de interesse).

Através deste indicador podemos obter informação acerca da existência ou não de similaridade entre as observações da variável de interesse - centraliza-se o valor de cada indicador local e, recorrendo ao diagrama de dispersão de *Moran*, obtêm-se para cada localização uma das quatro classes: alto-alto, baixo-alto, baixo-baixo e alto-baixo. Por exemplo, uma localização classificada na última classe significa que tem um valor alto da variável de interesse e que está rodeada por localizações com um valor baixo da mesma.

Para reforçar a medida anterior, devido ao facto deste teste não paramétrico ser pouco potente, utilizou-se a estatística G_i que permite obter a indicação, no caso de similaridade de

observações, se as zonas são similares devido a valores elevados ou pequenos. Esta medida é dada pela seguinte fórmula:

$$G_i = \frac{\sum_{j=1, j\neq i}^{n} w_{ij} x_j}{\sum_{j=1, j\neq i}^{n} x_j},$$

com as suposições análogas ao I local de *Moran* indicadas anteriormente. Um valor positivo indica um agrupamento espacial de valores altos da variável em estudo e um valor negativo desta estatística indica um agrupamento espacial de valores baixos.

De um modo geral, através da combinação das medidas indicadas é possível obter uma informação mais segura da distribuição geográfica da variável em estudo. As mesmas podem ser usadas em simultâneo uma vez que os pesos aplicados em ambas são iguais (Getis e Ord, 1992).

3.5 Análise de Agrupamentos

A análise de agrupamentos, mencionada no tópico 2.4, é composta por uma variedade de algoritmos e métodos de agrupamento com características específicas que os distingue dos restantes. Deste modo, a fase importante desta análise passa pela decisão do método de agrupamento que melhor se enquadra no conjunto de dados em estudo e tendo em conta o objetivo do mesmo. Para apoiar esta decisão seguiu-se os critérios: a forma como os grupos são formados; a estrutura dos dados; e a sensibilidade da técnica de agrupamento a alterações que não afetem a estrutura dos dados em estudo (Jain, Murty e Flynn, 1999; Zhao e Karypis, 2002; Everitt et al., 2011).

Com o objetivo de obter uma partição dos pacientes em estudo, escolheu-se um conjunto de métodos de agrupamento, segundo os critérios referidos anteriormente, que transmitissem segurança pelos resultados obtidos (consultar o Apêndice C para ver o código R respetivo).

Seguidamente apresenta-se as opções consideradas, de acordo com outros estudos encontrados na literatura (Hirano, Sun e Tsumoto, 2004; Everitt et al., 2011), para o cálculo da distância entre pares de observações.

- **Medida *Gower*** - esta medida de semelhança é apropriada para conjuntos de dados compostos por tipos de variáveis diversificados. O valor de similaridade

entre duas observações, entre N observações e p variáveis é dado pela seguinte expressão:

$$S_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}},$$

onde s_{ijk} é a similaridade entre a i-ésima e a j-ésima observação da k-ésima variável, ou seja, se variável for categórica s_{ijk} toma valor um se ambas observações partilharem a mesma categoria e o valor zero caso contrário. Se a variável for quantitativa aplica-se a seguinte regra:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k},$$

e R_k é o intervalo de valores da k-ésima variável. Relativamente a w_{ijk}, este toma valor zero se o valor de pelo menos uma das variáveis for omisso e toma o valor um no caso contrário. Por fim, obtém-se a matriz de distâncias $(N \times N)$ entre todas as observações através de:

$$D_{ij} = \sqrt{1 - S_{ij}},$$

sendo D_{ij} a distância entre a i-ésima observação e a j-ésima observação.

- **Combinação Linear da Medida *Mahalanobis* e da Medida *Hamming*** - a medida de *Mahalanobis* é adequada para variáveis quantitativas enquanto a medida *Hamming* é adequada para variáveis categóricas. Se a k-ésima variável for quantitativa tem-se:

$$d_{M_{ij}} = \left\{ (x_i - x_j)^T \sum^{-1} (x_i - x_j) \right\}^{\frac{1}{2}},$$

em que Σ representa a matriz de covariâncias $(p \times p)$ e Σ^{-1} permite padronizar os dados. Se a k-ésima variável for categórica tem-se:

$$d_{H_{ij}} = \frac{1}{p_c} \sum_{k=1}^{p_c} s_{ijk} \,,$$

onde p_c é o número de variáveis categóricas e s_{ijk} recai sobre o caso anterior. Por fim, obtém-se a seguinte combinação linear, considerando p_q o número de variáveis quantitativas:

$$D_{ij} = \frac{p_q}{p} \times d_{M_{ij}} + \frac{p_c}{p} \times d_{H_{ij}} \,.$$

Assim, obtém-se a matriz de distâncias $(N \times N)$ entre todas as observações recorrendo à combinação de duas medidas.

Optou-se pela combinação de métodos de agrupamento com o objetivo de captar os pontos fortes de cada algoritmo. A estratégia consistiu-se, essencialmente, em aplicar métodos hierárquicos aglomerativos para obter o número de grupos subjacente ao conjunto de dados, e aplicar o método de particionamento *k-means*, de modo a encontrar a partição dos dados segundo o número de grupos encontrado no método anterior.

Método Hierárquico Aglomerativo

O conceito base do método hierárquico aglomerativo baseia-se na distância entre as observações. O algoritmo começa por assumir que cada observação é um agrupamento distinto, seguidamente começa a aglomerar as observações mais próximas por um certo critério de distância e finaliza o algoritmo quando todas as observações fizerem parte do mesmo grupo.

A literatura encontrada (Hirano, Sun e Tsumoto, 2004) indica o critério de distância *ward*, na análise de dados de saúde, como aquele que providencia uma melhor qualidade dos agrupamentos obtidos - uma baixa dissemelhança intragrupos e uma elevada dissemelhança intergrupos - todavia considerou-se alguns critérios de distância para reforçar-se o resultado encontrado na literatura (Everitt et al., 2011):

- **vizinho mais próximo** - dois agrupamentos são aglomerados tendo em conta a distância mínima entre duas observações de cada um dos agrupamentos;
- **vizinho mais longe** - dois agrupamentos são aglomerados tendo em conta a distância máxima entre duas observações de cada um dos agrupamentos.

- **média do Grupo** - dois agrupamentos são aglomerados tendo em conta a distância média entre todos os pares de observações de cada um dos agrupamentos;
- ***ward*** - Dois agrupamentos são aglomerados tendo em conta que a distância é definida pela diferença entre a soma dos quadrados das distâncias intragrupo considerando os dois agrupamentos separados e a soma dos quadrados das distâncias intragrupo considerando os dois agrupamentos aglomerados.

A Figura 3.4 ilustra as primeiras três medidas de distância apresentadas.

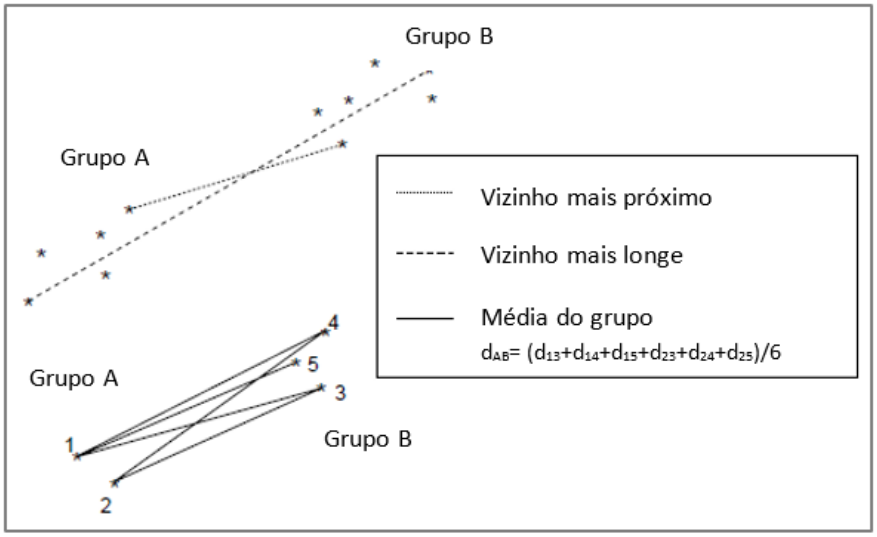

Figura 3.4 - Exemplo de três tipos de distância intergrupal (adaptado de Everitt et al., 2011)

Este método tem a particularidade de, uma vez alocada uma observação a um agrupamento, a mesma não poderá ser realocada, mesmo que o processo melhorasse a função de qualidade do método. Por outro lado, este tipo de método não requere à partida a indicação do número de grupos, sendo esta a sua principal vantagem em comparação com outros métodos.

O resultado do agrupamento encontrado apresenta-se na forma de um dendrograma, i.e., as observações apresentam-se agrupadas numa forma hierárquica. Assim, pode-se visualizar mais facilmente a aglomeração das observações e os grupos resultantes considerando um certo ponto de corte. Assim, uma vez que se desconhecia o ponto de corte mais indicado

neste conjunto de dados, considerou-se vários pontos obtendo-se, desta forma, um conjunto de números de grupos em cada agrupamento hierárquico aglomerativo considerado.

Considerou-se o coeficiente de agrupamento (ou semelhança intragrupo) como medida de qualidade dos agrupamentos obtidos. A partir deste coeficiente foi possível verificar qual das quatro medidas de distância descritas apresentava uma maior semelhança intragrupo, em cada um dos agrupamentos obtidos (Maechler et al., 2015).

Método de Particionamento

Dada a indicação do número de grupos, *k*, o algoritmo começa por selecionar *k* centros, designados por centróides. Seguidamente, cada observação é alocada ao centróide mais próximo, depois são calculados novos centróides e as observações são novamente realocadas. Como critério de paragem do algoritmo, considera-se a interação em que deixa de ser possível otimizar a função de qualidade do algoritmo.

Considerou-se os seguintes critérios de qualidade dos agrupamentos obtidos: a dissemelhança máxima e média entre as observações do grupo e o centróide do grupo; o diâmetro do grupo; o nível de separação do grupo; e a informação da silhueta de cada observação.

A silhueta de cada observação pretende quantificar até que ponto cada observação foi alocada ao grupo mais adequado (Rousseeuw, 1987; Maechler et al., 2015). Este indicador assume valores entre -1 e 1, sendo que um valor próximo da unidade indica que a observação foi alocada ao grupo mais adequado à sua silhueta, um valor positivo próximo de zero indica que a observação enquadra-se de forma análoga em mais do que um grupo, e um valor negativo indica que a observação poderá ter sido alocada ao grupo menos adequado a si segundo a sua silhueta.

4 RESULTADOS

Este capítulo centra-se na apresentação dos resultados obtidos através da metodologia descrita no capítulo anterior. Em primeiro lugar, apontam-se algumas características da população em estudo, de modo a analisar os resultados obtidos com algum conhecimento prévio da estrutura da população; seguidamente, é feita uma descrição da amostra e estatísticas descritivas da mesma; e, finalmente, faz-se uma apresentação dos resultados de cada uma das tarefas de inferência estatística, análise de dados espaciais e análise de agrupamentos.

4.1 CARACTERIZAÇÃO DEMOGRÁFICA DA POPULAÇÃO EM ESTUDO

Uma perceção geral da demografia da população residente na região ROR-Sul pode contribuir para uma melhor interpretação dos resultados obtidos. Assim, apresentam-se de seguida alguns pontos relativos à demografia da população em estudo.

Em relação aos residentes na região ROR-Sul, em 2013, a região de LVT agregava a maior proporção da população e a RAM agregava a menor. Observou-se também que em todas as regiões a população feminina predominava em relação à masculina (como se pode verificar na Tabela 4.1), e no total a feminina tinha uma ponderação de 53,2% do total da população.

Tabela 4.1 - População adulta residente nas regiões abrangidas pelo ROR-Sul, por sexo, a 30 de junho de 2013

Região	Masculino	Feminino	Total
Alentejo	178426	194234	372680
Algarve	179191	195668	374858
LVT	1463504	1675964	3139446
RAM	101312	119189	220505
ROR-Sul	1922433	2185055	4107489

Nota: adaptado de INE, 2015a

Relativamente à estrutura etária da população residente na região sul de Portugal em 2013, de um modo geral, não existiam diferenças aparentes entre as diversas regiões, como se pode verificar na Figura 4.1. Salienta-se que o grupo de indivíduos com idade igual ou superior a 75 anos tinha uma maior representatividade na região do Alentejo e menor na RAM (INE, 2015a).

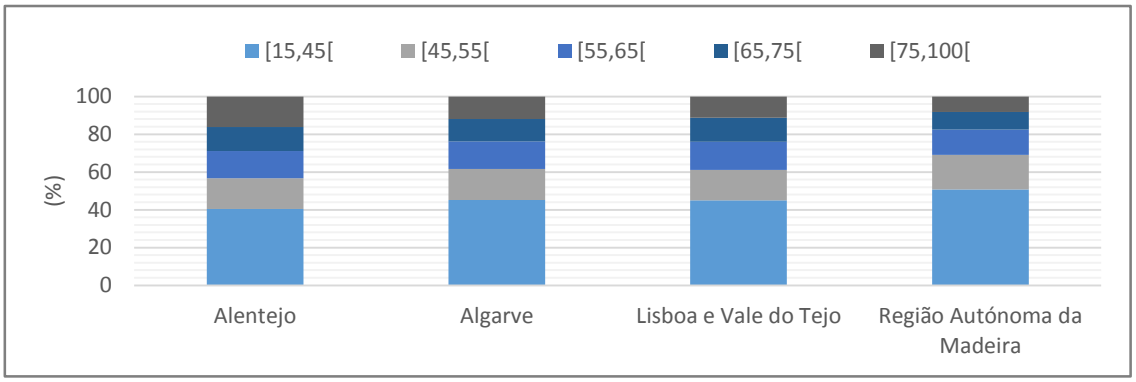

Figura 4.1- Estrutura etária da população, por região abrangida pelo ROR-Sul, a 30 de junho de 2013 (adaptado de INE, 2015a)

A estrutura etária desta população, por sexo e em 2013, assemelhava-se à estrutura etária de regiões desenvolvidas onde existe, essencialmente, uma tendência de envelhecimento demográfico devido ao acréscimo da esperança média de vida à nascença e ao decréscimo contínuo dos níveis de natalidade. De um modo geral, a população masculina predominava em relação à feminina aproximadamente até aos 50 anos, depois a tendência invertia-se, obtendo-se uma maior percentagem de mulheres idosas em relação à percentagem de homens idosos (ROR-Sul, 2014).

4.2 AMOSTRA

A amostra (*N*=958), de conveniência, é constituída por todos os casos de tumor maligno primário no pulmão, diagnosticados com ou sem confirmação de exames microscópicos, no primeiro semestre de 2013, em indivíduos residentes na região abrangida pelo ROR-Sul com idade igual ou superior a 15 anos.

A restrição da idade seguiu a nomenclatura de investigação usada pelo ROR-Sul, que realiza estudos segmentados entre pacientes adultos e infantis (ROR-Sul, 2014), tendo sido esta amostra sugerida pelos seus especialistas como de particular interesse, por ser representativa dos múltiplos desafios que o ROR-Sul enfrenta ao nível da análise de dados oncológicos.

4.3 ESTATÍSTICAS DESCRITIVAS

Neste tópico apresentam-se algumas estatísticas descritivas do conjunto de dados relativo aos RES, extraídos da base de dados do ROR-Sul. Este conjunto engloba 958 novos casos de cancro do pulmão, dos quais 22% são do sexo feminino e 78% são do sexo masculino, diagnosticados no primeiro semestre de 2013.

À data do diagnóstico, a média das idades era de 65,92 anos (desvio padrão (*s*) = 11,09 anos) e a mediana 66 anos. No sexo masculino a idade mínima era de 35 anos e a idade máxima era de 89 anos, a média de 66,08 anos (*s* = 10,61 anos) e a mediana 66 anos. No sexo feminino a idade mínima era de 22 anos e a idade máxima era de 91 anos, a média de 65,32 anos (*s* = 12,68 anos) e a mediana de 65,5 anos.

Pode-se observar na Figura 4.2 que a distribuição da idade é semelhante em ambos os sexos tendo em conta que o número de pacientes do sexo masculino é, sensivelmente, três vezes maior que o número de pacientes do sexo feminino. No entanto existe uma maior amplitude das idades dos pacientes do sexo feminino relativamente às idades dos pacientes do sexo masculino.

Verificou-se que o maior número de novos casos ocorreu em indivíduos com idade igual ou superior a 55 anos e que a tendência do aumento do número de casos com o aumento da idade verificou-se homologamente em ambos os sexos. Contudo, observou-se um número inferior de novos casos no grupo etário compreendido entre os 75 e 100 anos, no sexo masculino. Evidencia-se que a esperança média de vida aos 65 anos é de aproximadamente 17 anos para o sexo masculino e cerca de 20 anos para o sexo feminino (INE, 2015c).

Seguidamente, na Tabela 4.2, apresentam-se as classes de cada variável do conjunto de dados, referido anteriormente, onde se observou uma maior e menor frequência absoluta. Destaca-se que as informações omissas na Tabela 4.2 indicam que houve mais do que uma classe nessa situação.

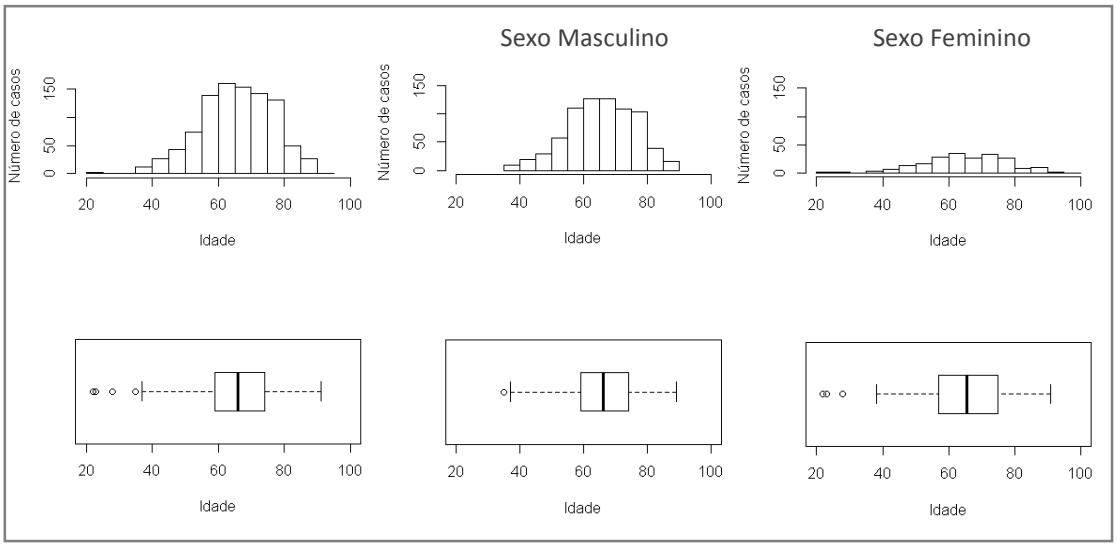

Figura 4.2 - Diagrama de extremos e quartis e histograma da idade dos pacientes, global e por sexo

Tabela 4.2 - Estatísticas descritivas de algumas variáveis categóricas em estudo

Descrição da Variável	**Classe com maior frequência**		**Classe com menor frequência**	
	Descrição da classe	Frequência absoluta	Descrição da classe	Frequência absoluta
Estado de vida	"Falecido"	562 (59%)	"Vivo"	396 (41%)
Topografia	"Pulmão SOE"	361 (38%)	"Múltiplas subcategorias do pulmão"	17 (2%)
Morfologia	"Adenocarcinoma SOE"	394 (41%)		
Grupos histológicos	"Adenocarcinomas"	400 (42%)	"Outros tipos de cancro específicos"	1 (0,1%)
Grau de diferenciação	"Desconhecido"	596 (62%)	"Indiferenciado"	16 (2%)
Estádio na apresentação	"Doença metastática"	435 (45%)	"Não aplicável"	29 (3%)

Nota: em parênteses encontra-se a respetiva frequência relativa

Quanto ao local de residência do paciente à data do diagnóstico, pode-se observar na Figura 4.3 que Lisboa é a localização com maior incidência. Destaca-se que não existiram registos de diagnósticos na Ilha de Porto Santo, em ambos os sexos, no primeiro semestre de 2013. A Ilha de Porto Santo tem 4 527 habitantes (dos quais 869 (19%) têm entre 55 a 74 anos), representando 11% da população em estudo na região ROR-Sul. Salienta-se ainda que no distrito de Portalegre não existe registo de incidência no que remete ao sexo feminino.

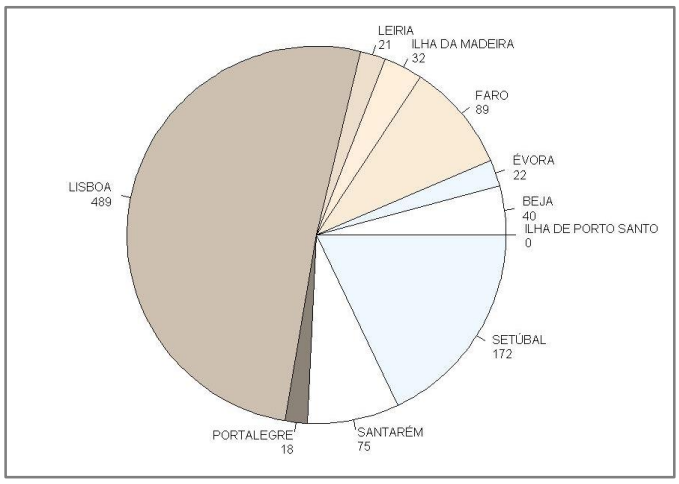

Figura 4.3 - Incidência do cancro do pulmão segundo o distrito ou ilha de residência no momento do diagnóstico

Ao verificar a percentagem de incidência ao longo das regiões do ROR-Sul concluiu-se que a região de LVT engloba 80%, seguida pela região do Algarve com 9%, a região do Alentejo abrange 8% e por fim a RAM com 3% da incidência na região ROR-Sul

De um modo geral, verificou-se que os concelhos com maior taxa de incidência situam-se na região do Alentejo (consultar o apêndice D para observar a distribuição das taxas de incidência ao longo dos concelhos da região ROR-Sul). O concelho de Alvito (distrito de Beja) verificou a maior taxa de incidência bruta do sexo masculino (186,20 por 100 000 habitantes) e o concelho de Monchique (distrito de Faro) a maior taxa de no sexo feminino (38,14 por 100 000 habitantes). Ainda em relação ao sexo feminino, verificou-se um número considerável de concelhos sem incidência, existindo em contrapartida, uma taxa significativa em concelhos isolados.

Em relação aos fatores ambientais (consultar o apêndice E para observar a sua distribuição ao longo dos concelhos da região ROR-Sul) verificaram-se diferenças ao longo dos diversos concelhos abrangidos pelo ROR-Sul, nomeadamente, o volume médio anual de precipitação foi maior na zona em redor da capital portuguesa e o número médio anual de incêndios teve uma maior presença no concelho de Sintra.

Relativamente ao IQA, uma vez que não existia informação da classificação para todos os dias do ano como referido anteriormente, não se pode concluir imediatamente que a zona a noroeste da região ROR-Sul tenha tido uma melhor qualidade do ar. Além disso, essa zona foi classificada com uma das classes menos favoráveis do IQA um número significativo de dias do ano. Observou-se, ainda, que os concelhos com um maior número médio de dias com essa classificação se situam nas zonas industriais em redor da capital portuguesa. Por fim, verificou-se que a qualidade do ar, segundo a média ponderada calculada, é pior na zona litoral e na região de LVT.

4.4 Inferência Estatística

Neste tópico apresentam-se os resultados obtidos pelo estudo de inferência estatística referido no tópico 3.3, nas três vertentes indicadas: comparação de valores médios de populações independentes; independência em tabelas de contingência; e regressão.

Comparação de Valores Médios de Populações Independentes

O objetivo desta tarefa consistia em testar a igualdade do valor médio das idades dos pacientes do sexo masculino e do sexo feminino, apesar do desconhecimento da distribuição das duas populações. Assim, começou-se por aplicar o gráfico Q-Q para comparar os quantis da distribuição empírica da idade dos pacientes no momento do diagnóstico com os da distribuição Normal (consultar o apêndice F para observar os gráficos Q-Q), a partir do qual se verificou que, nas extremidades do gráfico, a distribuição das idades dos pacientes afasta-se da Normal. Em relação à distribuição das idades de cada sexo observou-se que os quantis da distribuição de ambos os sexos se afastam da Normal, à semelhança do caso anterior.

De forma a confirmar as evidências encontradas nos gráficos Q-Q, aplicou-se o teste *Shapiro-Wilk* para a variável idade, a cada subpopulação: pacientes do sexo masculino e pacientes do sexo feminino. Dos resultados obtidos, rejeitou-se a hipótese nula nos dois casos (p <0,01). Portanto, existe evidência para afirmar que os dados não provêm de populações com distribuição Normal.

Uma vez que a normalidade dos dados dos dois casos foi rejeitada, testou-se se as populações por sexo seguiam a mesma distribuição no que diz respeito à idade. Com base no teste *Kolmogorov-Smirnov* para duas amostras, não se rejeitou a hipótese nula de igualdade das distribuições (p = 0,6715). Assim, não existe evidência para afirmar que as duas populações não provêm da mesma distribuição.

Independência em Tabelas de Contingência

Em primeiro lugar, obtiveram-se as tabelas de contingência através do cruzamento das variáveis relativas ao paciente (e.g.: sexo, estado de vida), com as variáveis relativas ao tumor (e.g.: sublocalização do tumor, estádio), com o intuito de evidenciar uma possível dependência entre as mesmas. Destaca-se que nas variáveis relativas ao tumor, as classes "desconhecido" e "não aplicável" não foram consideradas. Seguidamente, aplicaram-se os testes de independência indicados no tópico 3.3. Deste modo, a Tabela 4.3 apresenta os casos em que a hipótese nula foi rejeitada (resultantes do teste do *Qui-Quadrado*), ou seja, em que se evidenciou dependência entre os pares de variáveis (p <0,01).

Tabela 4.3 - Resultados significativos da análise de tabelas de contingência

Hipóteses	**Decisão**	**Conclusão**
H_0: estádio e distrito são independentes vs. H_A: estádio e distrito são dependentes	Rejeita-se H_0 (p <0,01)	Existe evidência para afirmar que o est**á**dio e o distrito de residência são dependentes
H_0: estado e distrito são independentes vs. H_A: estado e distrito são dependentes		Existe evidência para afirmar que o estado de vida e o distrito de residência são dependentes
H_0: estado e região são independentes vs. H_A: estado e região são dependentes		Existe evidência para afirmar que o estado de vida e a região de residência são dependentes

Observa-se, a partir da Tabela 4.3, que o estádio na apresentação e o estado de vida são dependentes do local de residência do paciente.

Regressão

O objetivo desta tarefa consistia no estudo de uma possível relação entre a taxa de incidência bruta (variável-resposta), os fatores ambientais (e.g.: *MedIncendios* e *IQArMelhor)* e a idade média dos pacientes, por concelho. Para este efeito, procedeu-se à análise de regressão múltipla referida no tópico 3.3, onde se testaram as combinações possíveis a partir das variáveis enquadradas nesta análise. Perante os resultados obtidos (consultar o apêndice G para observar alguns resultados), concluiu-se que, de todos os modelos avaliados e tendo em conta os resultados obtidos a partir do teste *T-Student* (p <0,001), R^2 (0,4149 – o valor observado não é elevado mas será discutido mais à frente) e AIC (975,01), o modelo mais adequado ao conjunto de dados seria:

$$Tx\ Incidência = 15{,}266 - 0{,}064 \times IQArMelhor + 0{,}368 \times MedIdade.$$

Através deste resultado pode-se analisar o seguinte exemplo: caso se aumente em 10 dias o *IQArMelhor* (melhores categorias do IQA) e se fixe o valor de *MedIdade* (idade média dos pacientes), a taxa de incidência irá diminuir em 0,64 valores; por outro lado, fixando a *IQArMelhor* e aumentando em 10 unidades a *MedIdade,* a taxa de incidência bruta irá subir 3,68 valores. Por outras palavras, o aumento do número de dias de uma das classes do IQA (muito bom ou bom) ou uma média de idade dos pacientes, mais baixa, irá implicar a diminuição da taxa de incidência. Salienta-se que a interação entre as duas variáveis referidas foi considerada, não se revelando significativa, o que significa que a relação entre cada variável-regressora e a variável-resposta se mantém, fixando os valores tomados pela outra variável-regressora.

Relativamente aos resíduos obtidos, d modelo apresentado anteriormente, observou-se que alguns pontos saem fora da linha reta, pelo que poderão ser simplesmente valores atípicos (consultar o apêndice G para observar os gráficos). Uma vez que não se pretende fazer predição com base neste modelo, e como a amostra é grande, o facto dos resíduos não se comportarem, no geral, de acordo com a distribuição Normal, não invalida os resultados obtidos.

Ainda no âmbito da análise de regressão, pretendia-se descrever o estádio do tumor em função da idade do paciente e dos fatores ambientais do concelho de residência do paciente. Dos modelos avaliados selecionou-se o modelo seguinte (p <0,001; AIC = 1014,1; Resíduo da Desviância (1006,1) muito elevado tendo em conta o número de graus de liberdade (745), mas, tal como no modelo anterior, a discussão será feita mais à frente):

$$Estádio = -8{,}006 + 0{,}134 \times Idade + 0{,}162 \times IQAr - 0{,}003 \times Idade \times IQAr.$$

Relembra-se que a variável-resposta *Estadio* é binária (0 – "doença local ou loco-regional", 1 – "doença metastática"). Assim, conclui-se que na presença de interação o efeito de uma das variáveis-regressoras na variável-resposta é diferente para diferentes valores da outra variável-regressora. Em termos práticos isto significa, por exemplo, que aumentando a idade do paciente em um ano e considerando um valor de *IQAr* elevado, a expressão: $0{,}134 - 0{,}003 \times IQAr$, irá tornar o valor da variável-resposta mais próximo de zero.

4.5 Análise de Dados Espaciais

Aplicaram-se os métodos de autocorrelação e associação espacial a todas as taxas de incidência calculadas nos seguintes conjuntos de dados espaciais pertencentes à região ROR-Sul: ROR-Sul continental, LVT, Alentejo e Algarve. Estas três subdivisões surgiram por duas razões: para evitar que algumas zonas dominantes impedissem outras de se realçar; e por se ter verificado diferenças entre as regiões em estudo. As duas razões indicadas estão relacionadas na medida que uma grande amplitude de valores (e.g.: região ROR-Sul continental) de uma variável de interesse implicará que, no teste de hipóteses adjacente, se exclua zonas menos significativas em relação a outras (e.g.: zonas da região LVT em comparação com zonas da região do Alentejo). Deste modo, depois de uma análise global, aconselha-se uma análise mais focada de forma a encontrar evidências não detetadas na primeira.

Optou-se por não se aplicar a toda a região ROR-Sul, e em particular à RAM, uma vez que foram mencionados na literatura alguns problemas na aplicação desta análise a um conjunto de dados espaciais composto por ilhas. Neste tipo de situações existem zonas sem vizinhos associados ou com um número muito reduzido, que iria interferir com a sensibilidade desta análise (Anselin et al., 2006; Bivand et al., 2008; Vidyattama, 2014).

Apresentam-se na Figura 4.4 dois resultados obtidos desta análise, nomeadamente as imagens de quartis da taxa de incidência bruta por sexo.

As imagens de quartis apresentadas na Figura 4.4 estão delimitadas pelos seguintes valores: valor mínimo, primeiro quartil, mediana, terceiro quartil e valor máximo da variável de interesse. O conjunto das cores: azul e cor-de-rosa representam a caixa do diagrama de extremos e quartis, ou seja, através destas imagens consegue-se perceber a dispersão das observações em cada variável. Através do teste Global de *Moran* verificou-se que em todas as

áreas geográficas aqui apresentadas não se rejeita a hipótese nula, i.e., não existem evidências para afirmar que não existe autocorrelação em cada uma das variáveis representadas na Figura 4.4.

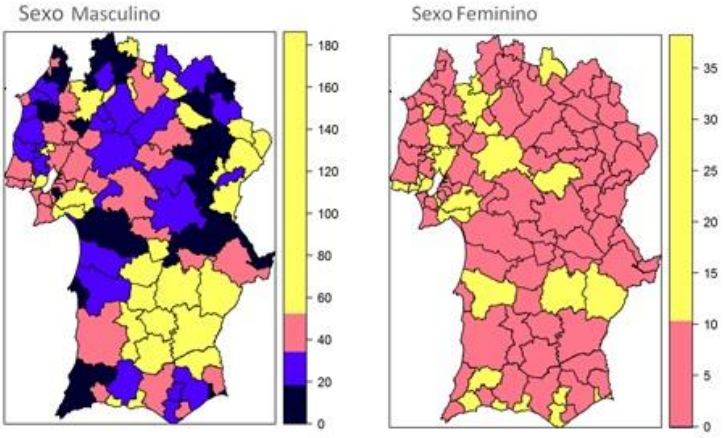

Figura 4.4 - Quartis da taxa de incidência bruta do sexo masculino e feminino, respetivamente, por concelho, em 2013

De um modo geral, para a área geográfica de Portugal continental, em 82% dos testes realizados não se rejeitou a hipótese nula e em 4% das variáveis existiu evidência de presença de autocorrelação espacial.

Dado que não se rejeitou a hipótese nula para as variáveis referidas, em particular neste tópico, foi possível realizar-se a análise de associação espacial para as mesmas. A Figura 4.5 apresenta os resultados obtidos para cada uma das variáveis mencionadas. Por observação das imagens verifica-se que apesar da ausência de um padrão global a análise local identificou agrupamentos de concelhos em algumas das classes de associação espacial local.

A Figura 4.6 apresenta as frequências absolutas de ocorrência dos concelhos, ao longo das sessenta e três taxas de incidência brutas e específicas para o caso de semelhança (classes: alto-alto e baixo-baixo) e dissemelhança (classes: alto-baixo e baixo-alto). Pode-se verificar que os concelhos das classes de semelhança detetados ao longo das diversas variáveis, apesar de um pouco dispersos, estão mais concentrados na região do Ribatejo e Alentejo, onde o concelho de Aljustrel (distrito de Beja) foi detetado mais vezes na classe de semelhança. Por outro lado, no que toca aos concelhos dissemelhantes a sua variabilidade é menor, apresentando

ocorrências na zona sul da Costa Vicentina, na zona central do distrito de Portalegre e na zona sudeste do distrito de Évora.

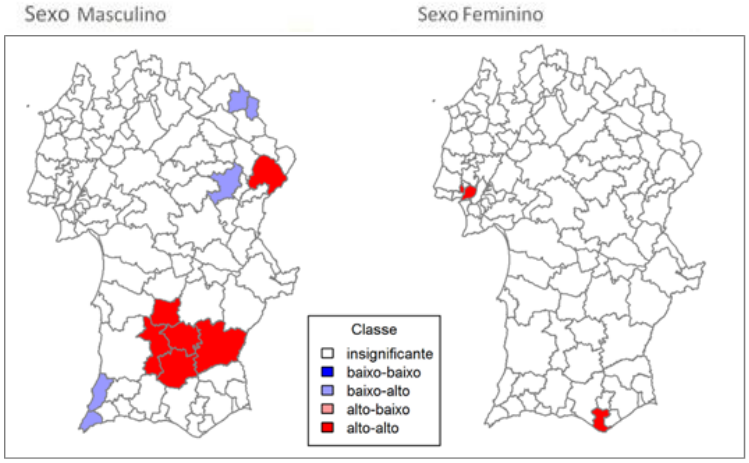

Figura 4.5 - Agrupamentos obtidos das variáveis correspondentes à taxa de incidência bruta do sexo masculino e feminino, respetivamente (p <0,05)

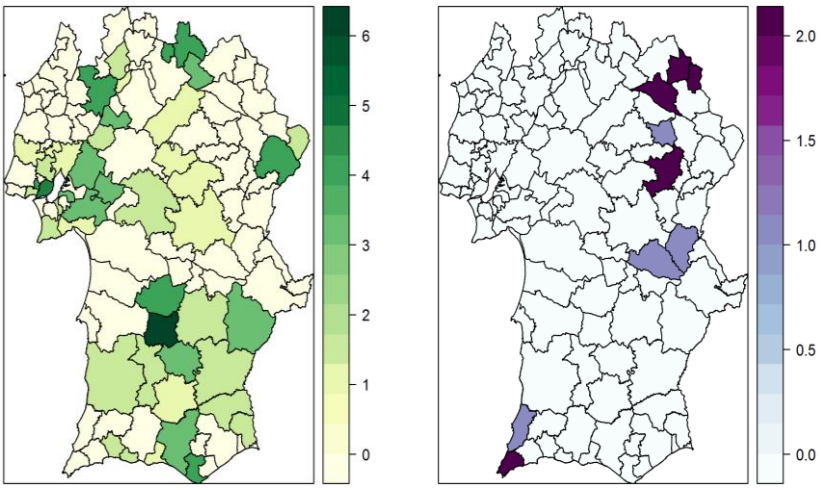

Figura 4.6 - Frequências absolutas de ocorrência dos concelhos por semelhança e dissemelhança, respetivamente

Por fim, salienta-se que a análise de dados espaciais apresentada teve, num computador pessoal, um tempo de computação em média inferior a cinco minutos[3], o que pode traduzir-se em um processo efetivamente eficiente para processamento de conjuntos de dados espaciais.

4.6 ANÁLISE DE AGRUPAMENTOS

Neste tópico apresentam-se os resultados obtidos pela aplicação de métodos de agrupamento ao conjunto de dados composto pelos RES dos pacientes e pelos fatores ambientais.

Em primeiro lugar, procedeu-se à seleção das variáveis que seriam incluídas nesta análise. Pretendia-se incluir variáveis relativas ao paciente, tumor e fatores ambientais; deste modo, numa primeira instância foram selecionadas as seguintes variáveis:

- sexo;
- idade no diagnóstico;
- sublocalização;
- grau de diferenciação;
- estádio;
- tipo de tumor;
- média de precipitação;
- média de incêndios.

Salienta-se que se escolheu a variável correspondente à idade do paciente em detrimento da faixa etária e do grupo etário, uma vez que as variáveis quantitativas conseguem obter uma medida de semelhança mais próxima da realidade.

Em relação às variáveis relativas à residência do paciente no diagnóstico (região ou distrito) e à qualidade do ar (*IQAr, IQArMelhor ou IQArPior*) mais adequadas procedeu-se a vários testes, tanto a nível do conjunto de dados de entrada como de várias medidas de modo a encontrar a melhor estratégia para esta análise. A Tabela 4.4 apresenta um sumário de alguns testes realizados com recurso a métodos de agrupamento hierárquicos, onde se pode verificar que para todos os conjuntos de variáveis e medidas de distância o método de agrupamento

[3] Considerando um processador Intel® Core™ i3 CPU M350 @ 2.27GHz e uma memória RAM de 4 GB (3,79 GB usáveis)

ward assume um valor de coeficiente de agrupamento (semelhança intragrupo) mais elevado que os restantes métodos.

A partir dos resultados apresentados na Tabela 4.4, pode-se concluir que o coeficiente de agrupamento mais elevado (≈ 0.9932) foi obtido a partir das variáveis-base, apresentadas anteriormente neste tópico, e as variáveis *IQAr* e *RegiaoDiag*, considerando a medida *gower* para o cálculo da matriz de distâncias entre as observações e aplicando o método *ward* para o processo de constituição dos agrupamentos.

A combinação destes métodos teve também o melhor desempenho para o conjunto de variáveis-base incluindo a *RegiaoDiag*, *IQArMelhor* e *IQArPior*, com um coeficiente próximo do que foi referido anteriormente (≈ 0,9930). As análises posteriores focaram-se nos resultados com coeficiente superior, tendo sido verificado, no entanto, um desempenho elevado para todos os conjuntos de variáveis utilizando estes métodos.

Tabela 4.4 - Descrição da medida de qualidade, o coeficiente do agrupamento, dos agrupamentos hierárquicos considerados

Variáveis em Teste		**Medida**	**Métodos de Agrupamento dos Dados**			
			Média do grupo	Vizinho mais próximo	Vizinho mais longe	*Ward*
IQAr	*RegiaoDiag*	*Gower*	0,908550618	0,819539362	0,937294173	**0,993286**
	RegiaoDiag	*M e H*	0,838185578	0,775087352	0,916240066	**0,984466**
	DistritoDiag	*Gower*	0,887740658	0,791975733	0,923766766	**0,991284**
	DistritoDiag	*M e H*	0,857624819	0,81531563	0,92649937	**0,986494**
IQArMelhor e *IQArPior*	*RegiaoDiag*	*Gower*	0,907945011	0,809561505	0,933846686	**0,993063**
	RegiaoDiag	*M e H*	0,905719208	0,881280669	0,945481629	**0,988637**
	DistritoDiag	*Gower*	0,892396487	0,781307649	0,923329329	**0,991595**
	DistritoDiag	*M e H*	0,910595167	0,88979888	0,94844312	**0,989263**

Nota: "M e H" indica a combinação das medidas Mahalanobis e Hamming

Os números de grupos: 3, 10, 17 e 21, resultaram do corte do dendrograma em diferentes alturas (consultar o apêndice H para observar o dendrograma obtido), obtido pelo método *ward* nas condições descritas anteriormente. Posteriormente, todos os números de grupos encontrados foram aplicados ao método de particionamento dos dados, *k-means*. Dada

a natureza exploratório deste trabalho, não existia indicação de número de grupos desejável, pelo que estes são apresentados como possíveis conjuntos que podem eventualmente servir para uma análise clínica posterior.

Na Figura 4.7 apresenta-se um dos resultados obtidos (*k*=3) para se compreender a localização de cada observação no espaço, representando, desta forma, a sua dissimilaridade obtida a partir da matriz de distâncias (consultar o apêndice I para observar um outro resultado). Verifica-se a existência de uma separação não linear entre os diferentes grupos e por esse motivo a maioria das categorias de cada característica não ter uma ponderação significativamente superior às restantes em casa grupo. No entanto, verificou-se que uma certa separação a nível geográfico, nomeadamente, um dos grupos era composto, na sua maioria, por pacientes residentes no distrito de Évora, Faro e Portalegre.

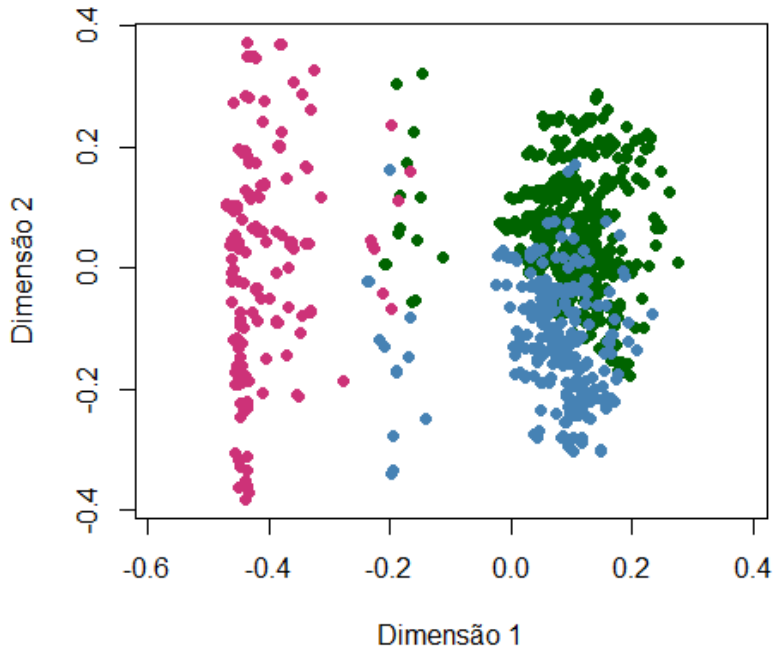

Figura 4.7 - Representação gráfica das observações considerando 3 grupos

O gráfico da Figura 4.8, denominado gráfico de silhueta, pode ser interpretado do seguinte modo: cada barra vertical representa uma observação que tem associada um valor de silhueta; se for próximo de um indica que a observação foi alocada ao grupo mais adequado a

si; se o valor for positivo mas próximo de zero significa que a observação pode-se adequar a mais do que um grupo; e se for negativo a observação poderá ter sido alocada a um grupo que não seria o mais adequado para si, ou seja, existe uma dissemelhança significativa entre a observação e as restantes observações do grupo (Rousseeuw, 1987).

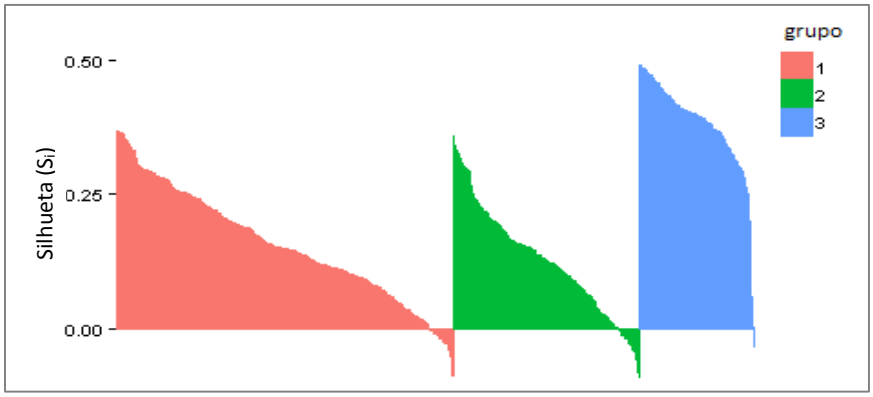

Figura 4.8 - Informação da silhueta das observações considerando 3 grupos

Na figura 4.8 verifica-se que os níveis de silhueta encontram-se próximos de zero o que significa, em termos práticos, que o conjunto de dados não admite uma separação dos dados linear. Ainda se pode constatar que existe um certo número de observações que possivelmente encontram-se afetas ao grupo menos adequado para si, segundo a sua semelhança com as restantes observações dentro do mesmo grupo.

Por fim, salienta-se que cada teste considerado nesta análise de agrupamentos teve, num computador pessoal, um tempo de computação em média inferior a oito minutos[4], o que pode ser considerado um tempo aceitável para análises deste género, onde a necessidade de resultados em tempo muito curto não é elevada.

[4] Considerando um processador Intel® Core™ i3 CPU M350 @ 2.27GHz e uma memória RAM de 4 GB (3,79 GB usáveis)

5 DISCUSSÃO DOS RESULTADOS

Tendo em conta a metodologia indicada no capítulo 3, este capítulo tem como finalidade apresentar os principais desafios encontrados, os resultados obtidos e a discussão dos mesmos.

Durante a extração, limpeza e integração dos dados deste estudo, verificaram-se alguns desafios, nomeadamente a inacessibilidade e estrutura da informação, a ausência de dados, a presença de valores omissos e heterogeneidade da codificação de dados utilizada pelas diversas fontes de dados que, por sua vez, podem ter influenciado os resultados obtidos.

Relativamente à informação presente nas fontes de dados disponíveis para domínio público, esta é apresentada em interfaces gráficas, dificultando a obtenção de dados em maior escala que possibilitem trabalhos mais abrangentes. Também a estrutura da informação, na sua origem, encontrava-se desadequada para a sua aplicação aos métodos utilizados neste estudo por exemplo - a informação referente ao IQA encontrava-se repartida em várias variáveis.

No que toca ao registo de dados, observou-se a ausência do mesmo nas características referentes ao paciente, nomeadamente os comportamentos fundamentais para se compreender melhor o fenómeno do cancro do pulmão, dado serem apontados pela literatura como fatores de risco para o seu desenvolvimento (e.g.: hábitos de alimentação, atividade física e tabágicos).

Ainda existe a questão da indisponibilidade de dados exaustivos, abrangentes e adequados a qualquer utilizador relativamente aos fatores ambientais a nível nacional, por exemplo - dados relativos à cobertura florestal, nível de poluição e níveis de radiação.

No tópico dos valores omissos, existe efetivamente na base de dados do ROR-Sul um conjunto de campos abrangente e exaustivo referente às características do tumor, como indicado no tópico 3.2.2. e no Anexo 2. Contudo, com a extração deste conjunto, na tentativa de especificar o tipo de tumor, obter-se-iam muitos valores omissos, não existindo uma forma eficiente de completar esta informação. Por outro lado, se se escolher um conjunto de características sem valores omissos, este seria menos detalhado, o que iria traduzir-se na possibilidade de perda de informação útil para o processamento dos dados.

Em relação aos dados relativos aos fatores ambientais, apesar da existência de alguma informação nas fontes de dados referidas neste estudo, estas apresentam um número

significativo de valores omissos. Este facto pode levar a interpretações pouco realistas do panorama ambiental atual a nível nacional, e em particular na região ROR-Sul.

Relativamente ao último desafio indicado, a heterogeneidade da codificação, está relacionada com o facto de fontes de dados diferentes considerarem diferentes codificações para as mesmas observações. Esta questão tem particular relevância na integração dos diferentes conjuntos de dados recolhidos de natureza geográfica. Apesar de existir uma codificação para cada localização geográfica em Portugal regularizada pela DGT, esta não é uniforme nas diferentes fontes usadas neste estudo. Como consequência, utilizou-se a designação de cada localização, estratégia possível devido à existência de designações únicas dentro da região ROR-Sul (que não tem designações de localizações repetidas). No entanto a nível nacional não seria possível aplicar esta estratégia por existirem designações de localizações repetidas e impossíveis de desambiguar.

Outros aspetos relacionados com a nomenclatura utilizada a nível geográfico pelas fontes de dados prendem-se com o facto de que cada fonte de dados utilizar delimitações territoriais diferentes, levando à necessidade de recorrer a uma integração quase manual, e ainda a identificação das respetivas delimitações territoriais para cada caso.

Não obstante estes desafios, foi possível analisar dados de 958 pacientes integrados com dados relativos à qualidade do ar e incêndios para todas as regiões geográficas em estudo. Deste modo, é possível concluir que se cumpriu o primeiro desafio (extração, limpeza e integração de dados de natureza diversa e investigação dos obstáculos), sendo este complementar ao objetivo principal.

Em relação à análise de dados espaciais, esta revelou ser útil para verificar a existência de padrões geográficos detetados nas diversas variáveis estudadas através das imagens geográficas obtidas. Este tipo de análise é de particular relevância para a aplicação em casos onde a comparação de zonas geográficas seja desejável, como é o caso do ROR-Sul, que tem como objetivo a análise de dados oncológicos ao nível regional. Contudo, esta análise não foi aplicada à Região Autónoma da Madeira devido às limitações da sua aplicação a ilhas, o que interferiria com a sensibilidade dos testes estatísticos. A aplicação da análise de dados espaciais foi bem sucedida, revelando diversos padrões geográficos de potencial interesse. Em particular, identificou-se um conjunto de concelhos pertencentes, na sua maioria, ao distrito de Beja, cuja incidência é superior em relação aos restantes concelhos. Também verificou-se no distrito de Portalegre a ocorrência de alguns concelhos cuja incidência apresenta valores atípicos em relação aos seus concelhos vizinhos. Desta forma, cumpriu-se o segundo desafio complementar

deste estudo referente à possível utilidade de métodos de análise de dados espaciais na identificação de padrões úteis.

Complementarmente à análise de dados espaciais, realizou-se um estudo de agrupamento de dados. Nesta etapa, foram encontrados vários desafios, tais como a seleção do conjunto de variáveis a utilizar, a definição de uma medida de distância, a seleção de algoritmos de agrupamento e a parametrização dos mesmos.

No que se respeita ao conjunto de dados desejáveis de serem utilizados nesta análise, constataram-se dois factos: a existência de uma lacuna na caracterização do paciente (número insuficiente de variáveis) e a presença de um número considerável de variáveis categóricas que captam, de uma forma menos precisa, a semelhança entre os pacientes.

Um outro aspeto refere-se à forma como os tipos de tumor são codificados, codificação essa que não permite uma comparação da sua semântica, sendo do conhecimento geral a existência de alguns tipos de tumor mais semelhantes entre si. Assim, nos métodos aplicados, cada tipo de tumor foi considerado igualmente dissimilar a todos os outros, o que não reflete a complexidade real do domínio em análise.

A metodologia utilizada combina um primeiro passo de agrupamento hierárquico para definição do número de agrupamentos, seguida de *k-means* com o número selecionado. Esta mostrou-se mais eficiente quando comparada com outras metodologias utilizadas noutros estudos baseados na aplicação de um único método de separação de dados, como por exemplo o de agrupamento hierárquico aglomerativo (Tabela 2.3).

À semelhança do que tinha sido indicado na literatura, o método de agrupamento *ward* provou ser mais eficiente em comparação com os restantes, uma vez que se verificou que o coeficiente do agrupamento mais elevado (≈ 0,993) tinha sido obtido através do método *ward* e da medida de *gower* para o cálculo da matriz de distâncias.

Salienta-se que este tipo de análise requer a definição de parâmetros por parte do investigador (e.g.: altura de corte do dendrograma), e ao substituir estas decisões manuais por outros métodos complementares, consegue-se garantir uma maior segurança dos resultados obtidos.

Deste modo, a combinação de vários métodos de agrupamento e o teste de vários argumentos de entrada para estes, traduziu-se numa maior confiabilidade, qualidade e solidez dos resultados obtidos.

De uma forma geral, seria esperado que os grupos obtidos fossem constituídos por observações homogéneas entre si e heterogéneas entre as de outros grupos. No entanto

a análise dos resultados obtidos demonstra que os grupos não cumprem esta expetativa, sendo as observações de um grupo pouco semelhantes entre si.

Pode-se concluir que a complexidade do fenómeno do cancro do pulmão não parece ser totalmente explicada pelas variáveis e modelos criados dado que, por exemplo, podem existir certos fatores que são ainda clinicamente desconhecidos como fatores de risco para o desenvolvimento do cancro. Por outro lado, encontraram-se resultados relevantes, nomeadamente algumas evidências ao nível geográfico, como por exemplo a presença de um grupo constituído maioritariamente por pacientes que residem na região do Alentejo e do Algarve.

A razão desta separação a nível geográfico poderá ter sido a presença predominante de um conjunto de características relacionadas com a residência do paciente (e.g.: fatores ambientais) ou outros motivos não captados pelas variáveis utilizadas nesta análise. Desta forma, foi possível atingir o objetivo principal deste estudo - extrair informação útil dos RES, como também os desafios complementares indicados.

Tanto na análise de agrupamentos como na análise de dados espaciais, criaram-se modelos que podem ser aplicados a outros conjuntos de dados com características semelhantes, nomeadamente para investigações de outras doenças. É igualmente importante referir que o tempo de computação é significativamente reduzido, permitindo uma análise global em tempo reduzido.

Da análise dos dados realizada verifica-se que as características da incidência do cancro do pulmão na região ROR-Sul no primeiro semestre de 2013 encontram-se próximas do que já é conhecido na literatura.

A incidência do cancro do pulmão foi superior em indivíduos do sexo masculino e numa faixa etária mais elevada. Em contrapartida, constatou-se que a população feminina é significativamente superior à masculina e que a mediana de idades dos pacientes está próxima da esperança média de vida. Outros factos preocupantes são, por exemplo, a existência de uma porção elevada de casos com sublocalização do tumor desconhecida e o crescente aumento de novos casos de adenocarcinomas. Por outras palavras, pode-se considerar que cada vez mais o cancro se desenvolve de forma mais invasiva e que se encontra em mutação em relação ao que é atualmente conhecido ainda que numa tendência pouco evidente.

A nível geográfico confirma-se a existência de diferenças, tanto no sexo dos pacientes, como na sua faixa etária. De um modo geral, a incidência do cancro do pulmão na região ROR-Sul, no primeiro semestre de 2013, teve uma maior presença na região do Alentejo. Da análise

de tabelas de contingência verificou-se que tanto a região como o distrito de residência do paciente estariam dependentes do estádio do tumor e do estado de vida dos pacientes. Desta forma, verifica-se que ao nível geográfico existe uma dependência relacionada com o estado de desenvolvimento do tumor, ao contrário do que foi verificado noutro estudo a nível continental (tópico 2.5.2).

Dos resultados obtidos da análise de regressão, a taxa de incidência do cancro do pulmão poderá ser uma resposta conjunta do nível de qualidade do ar e da média de idade dos pacientes por concelho. Verificou-se que quanto maior o número de dias do ano com uma boa classificação do ar e quanto menor for a média de idade dos pacientes no diagnóstico, menor será a taxa de incidência. Este resultado encontra-se de acordo com outros estudos semelhantes, referidos no tópico 2.3, onde se encontraram evidências do impacto da qualidade ambiental na incidência do cancro do pulmão. No entanto, apontou-se duas questões nesta análise: para o primeiro modelo, teria sido mais adequado aplicar à análise de regressão múltipla a idade média da população por concelho, em vez da idade média dos pacientes (contudo, esta informação não se encontra disponível), e ainda, apesar das medidas utilizadas indicarem à partida um fraco ajustamento dos modelos indicados em certos casos, as mesmas podem não ser coerentes, no sentido que, por exemplo, um $R^2 = 0,75$ (valor elevado) poderá não implicar associação linear forte (Pestana e Velosa, 2008). A explicação deste facto pode estar na existência de alguma interferência de outras variáveis, i.e., podem existir mais fatores associados à variável resposta que não tenham sido considerados (e.g.: poluição atmosférica automóvel e radão). Não se consideraram estes e outros fatores, uma vez que os mesmos não se encontram disponíveis publicamente.

Em suma, apesar dos crescentes esforços para compilar informação útil e relevante, estudos mais abrangentes e complexos como este demonstram a necessidade de uma recolha mais alargada de outros tipos de dados, que permitam contextualizar a informação típica e fundamentar análises mais aprofundadas.

Também se pode concluir que a incidência do cancro do pulmão tem uma tendência cada vez mais progressiva, o que associado às evidências significativas do impacto dos fatores ambientais na incidência do cancro do pulmão em alguns concelhos da região ROR-Sul, alerta para a necessidade de alargar o estudo da incidência deste tipo de cancro a fatores não-clínicos.

6 CONCLUSÃO

Uma das peças fundamentais neste tipo de investigação é a existência de informação que caracterize tanto quanto possível as observações em estudo. Nesta investigação pretendia-se incluir, entre outros, dados relativos aos comportamentos de risco individuais, como por exemplo os hábitos tabágicos do paciente, sinalizados como relevantes para o desenvolvimento de cancro.

Contudo, esta informação não faz parte do conjunto de itens de preenchimento obrigatório no registo oncológico. À semelhança do que foi referido, também os fatores ambientais apresentaram informações omissas e inconsistentes, afetando a qualidade do estudo e a correta interpretação dos resultados.

Como trabalho futuro seria interessante realizar uma análise semelhante mas com um conjunto de variáveis descritivas dos pacientes, do diagnóstico e do tumor mais abrangente e exaustivo. Além disso, este conjunto de variáveis teria que ser mais diversificado, no sentido de também caracterizar da melhor forma possível o quotidiano do paciente. Este aspeto é importante porque o cancro desenvolve-se ao longo de muitos anos e, assim sendo, todos os fatores envolvidos na vida do paciente (inclusive a sua profissão) deveriam ser considerados.

Na análise de agrupamentos, onde se testaram várias opções, constatou-se que os resultados das medidas de qualidade foram ao encontro da literatura e que a combinação de vários métodos providenciou uma maior segurança nos resultados obtidos. Verificou-se que o método de agrupamento dos dados, *ward,* foi o mais adequado, uma vez que o maior valor do coeficiente do agrupamento foi encontrado neste caso. Como trabalho futuro, relativamente ao ponto anterior, seria interessante uma maior investigação das alturas de corte possíveis no dendrograma e avaliar com outras medidas de qualidade complementares os agrupamentos obtidos, uma vez que a escolha dos argumentos dos métodos por parte do investigador torna a investigação mais suscetível a ir ao encontro do que já é do conhecimento geral.

Os modelos descritivos construídos permitem a sua aplicação noutros conjuntos de dados que mantenham a mesmo estrutura, ou seja, é possível aplicar estes algoritmos no estudo de outro tipo de cancro ou doença. O tempo de computação dos métodos é consideravelmente reduzido, em média inferior a cinco minutos, o que permite análises em temporalmente mais rápidas.

Em suma, esta investigação evidencia os desafios atuais na prospeção integrada de dados ambientais e clínicos oncológicos na região sul de Portugal. Considerando que o ROR-Sul é o registo oncológico mais completo do país e com maior cobertura a nível Europeu, este estudo é revelador dos importantes desafios que este tipo de estudo enfrenta.

Ainda assim, foi possível obter resultados relevantes que, doravante, e utilizando as metodologias aqui propostas, poderão orientar estudos de outros tipos de cancro.

LISTA DE REFERÊNCIAS

Agência Portuguesa do Ambiente (s.d.). QualAr - Índices. Consultado a 5 de Setembro de 2015 em http://qualar.apambiente.pt/

American Psychological Association. (2012). *APA style to electronic references* (6ª ed.). Washington, D. C.: Autor.

Anderson, J. G. (2007). Social, ethical and legal barriers to e-health. *International journal of medical informatics*, 76(5), 480-483.

Anselin, L. (1995). Local Indicators of Spatial Association. Ohio: University Press.

Anselin, L., Syabri, I., e Kho, Y. (2006). GeoDa: an introduction to spatial data analysis. *Geographical analysis, 38*(1), 5-22.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., ... e Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences, 98*(24), 13790-13795.

Biggeri, A., Barbone, F., Lagazio, C., Bovenzi, M., e Stanta, G. (1996). Air pollution and lung cancer in Trieste, Italy: spatial analysis of risk as a function of distance from sources. Environmental Health Perspectives, 104(7), 750.

Bivand, R., Pebesma, E. J., e Gómez-Rubio, V. (2008). Applied Spatial Data Analysis with R. New York: Springer.

Breault, J. L., Goodall, C. R., e Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine, 26*(1), 37-54.

CancerCare. (2015a). What is Lung Cancer? Consultado a 7 de Agosto de 2015 em http://www.lungcancer.org/find_information/publications/163-lung_cancer_101/265-what_is_lung_cancer

CancerCare. (2015b). Risks. Consultado a 7 de Agosto de 2015 em http://www.lungcancer.org/find_information/publications/163-lung_cancer_101/273-risks

Cody R., Ed D., Wood R., e Medical J. (s.d.). Data Cleaning 101 Variable Name Variable Type Valid Values.

de Jonge, E., e van der Loo, M. (2013). *An introduction to data cleaning with R*. Technical Report 201313, Statistics Netherlands, 2013. URL http://www.cbs.nl/nl-NL/menu/methoden/onderzoek-methoden/discussionpapers/archief/2013/default.htm.

Del Giglio, A. (1996). *Câncer: Introdução ao seu Estudo e Tratamento*. São Paulo: Pasqualin.

Denoix, P. (1977). *A Cancerologia*. Lisboa: Publicações Dom Quixote.

Direção Geral do Território. (2015). Carta Administrativa Oficial de Portugal em Vigor. Consultado a 10 de Fevereiro de 2015 em http://www.dgterritorio.pt/cartografia_e_geodesia/cartografia/carta_administrativa_oficial_de_portugal__caop_/caop_em_vigor/

Draper, N.R., e Smith, H. (1998). *Applied Regression Analysis* (3ª ed.). John Wiley and Sons.

Edge, S. B., Byrd, D. R., Compton, C. C., Fritz, A., Greene, F. L., e Trotti, A. (Ed.). (2010). *AJCC Cancer Staging Handbook* (7ª ed.). Chicago: American Joint Committee on Cancer.

Everitt, B., e Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.

Everitt, B., Landau, S., Leese, M., e Stahl, D. (2011). *Cluster Analysis* (5ª ed.). Wiley Series in Probability and Statistics.

Faraway, J.J. (2004). *Linear Models with R*. Chapman & Hall/CRC.

Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery in databases. *Al Magazine, 17*(3), 37-53.

Franceschini, J., Jardim, J. R., Fernandes, A. L. G., Jamnik, S., e Santoro, I. L. (2013). Relação entre a magnitude de sintomas e a qualidade de vida: análise de agrupamentos de pacientes com câncer de pulmão no Brasil. *J Bras Pneumol, 39*(1), 23-31.

Fritz, A., Percy, C. Jack, A., Shanmugaratnam K., Sobin, L., Parkin, D. M., e Whelan S. (Ed.).(2000). *International Classification of Diseases for Oncology* (3ª ed.). Geneva: World Health Organization.

Fundação Portuguesa do Pulmão. (s.d.). Cancro do Pulmão. Consultado a 7 de Agosto de 2015 em http://www.fundacaoportuguesadopulmao.org/cancro_do_pulmao.html

GeoDa Center. (s.d.). Software. Consultado em 15 de Fevereiro de 2015 em https://geodacenter.asu.edu/software

Getis, A., e Ord, K. (1992). *The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis* 24, 189-206.

Girard, L., Zöchbauer-Müller, S., Virmani, A. K., Gazdar, A. F., e Minna, J. D. (2000). Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering. *Cancer research, 60*(17), 4894-4906.

Goovaerts, P., e Jacquez, G. M. (2004). Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics, 3*(1), 14.

Hagar, Y., Albers, D., Pivovarov, R., Chase, H., Dukic, V., e Elhadad, N. (2014). Survival analysis with electronic health record data: Experiments with chronic kidney disease. Statistical Analysis and Data Mining: *The ASA Data Science Journal, 7*(5), 385-403.

Han, J., Kamber, M., e Pei, J. (2011). *Data Mining: Concepts and Techniques* (3ª ed.). San Francisco: Morgan Kaufmann Publishers.

Hand, D. J., Mannila, H., e Smyth, P. (2001). *Principles of data mining*. Massachusetts: MIT Press.

Hinojosa de la Garza, O. R., Sanín, L. H., Montero Cabrera, M. E., Serrano Ramirez, K. I., Martínez Meyer, E., e Reyes Cortés, M. (2014). Lung Cancer Mortality and Radon Concentration in a Chronically Exposed Neighborhood in Chihuahua, Mexico: A Geospatial Analysis. The Scientific World Journal, 2014.

Hirano, S., Sun, X., e Tsumoto, S. (2004). Comparison of clustering methods for clinical databases. Information Sciences, 159(3), 155-165.

Iakovidis, I. (1998). Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in Europe. *International journal of medical informatics, 52*(1), 105-115.

Instituto Nacional de Estatística. (2015a). População residente (N.º) por Local de residência (NUTS - 2013), Sexo e Grupo etário; Anual. Consultado a 1 de Julho de 2015 em https://www.ine.pt/xportal/xmain?xpid=INEexpgid=ine_indicadoresecontecto=pieindOcorrCod=0008273eselTab=tab0

Instituto Nacional de Estatística. (2015b). Incêndios florestais (N.º) por Localização geográfica (NUTS - 2002); Anual. Consultado a 30 de Setembro de 2015 em https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0001145&contexto=pi&selTab=tab0

Instituto Nacional de Estatística. (2015c). Esperança de vida aos 65 anos (Metodologia 2007 – Ano) por Sexo; Anual. Consultado a 15 de Setembro de 2015 em https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0001723&contexto=bd&selTab=tab2

Instituto Nacional de Saúde Dr. Ricardo Jorge. (2010). Inquérito Nacional de Saúde 2005-2006. Consultado a 30 de Setembro de 2015 em http://www.insa.pt/sites/INSA/Portugues/Publicacoes/Outros/Paginas/INS2005-2006.aspx

International Agency for Research on Cancer. (2015). GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012 [Fact Sheets]. Consultado a 6 de Agosto de 2015 de http://globocan.iarc.fr/Pages/fact_sheets_population.aspx

Ismael, G. F. V., Coradazzi, A. L., Neto, F. A., Abdalla, K. C., Milhomem, P. M., Oliveira, J. D. S., ... e Segalla, J. G. M. (2010). Aspectos clínicos e histopatológicos em câncer de pulmão: análise dos dados de uma instituição no interior paulista entre 1997 e 2008. Revista Brasileira de Oncologia Clínica Vol, 7(22).

Jain, A. K., Murty, M. N., e Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264-323.

Jensen, P. B., Jensen, L. J., e Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405.

Krishnan, V. G., Ebert, P. J., Ting, J. C., Lim, E., Wong, S. S., Teo, A. S., ... e Ng, P. C. (2014). Whole-genome sequencing of Asian lung cancers: second-hand smoke unlikely to be responsible for higher incidence of lung cancer among Asian never-smokers. *Cancer research, 74*(21), 6071-6081.

Kristianson, K. J., Ljunggren, H., e Gustafsson, L. L. (2009). Data extraction from a semi-structured electronic medical record system for outpatients: a model to facilitate the access and use of data for quality control and research. *Health informatics journal, 15*(4), 305-319.

Lin, X. L., Chen, Y., Gong, W. W., Wu, Z. F., Zou, B. B., Zhao, J. S., ... e Jiang, J. M. (2013). Geographic distribution and epidemiology of lung cancer during 2011 in zhejiang province of china. Asian Pacific journal of cancer prevention: APJCP, 15(13), 5299-5303.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., e Darnell, J. (2000). *Molecular Cell Biology* (4ª ed.). York, United Kingdom: W. H. Freenan and Company.

Lunet, N., e Pimentel, P. (2012). Registo Oncológico de Base Populacional em Portugal: Reflexão sobre a Situação Atual e Perspetivas Futuras, 124-128.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., e Roudier, P. (2015). Package 'cluster'.

Mitchell, T. M. (1999). Machine Learning and Data Mining Over the past, *42*(11).

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Miller, W. G., ... e Knaus, W. A. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers in biology and medicine, 36*(12), 1351-1377.

Murteira, B., e Antunes, M. (2012). *Probabilidades e Estatística* (vol. II). Lisboa: Escolar Editora.

Needham, R. M. (1965). Classification and Grouping: Computer methods for classification and grouping. In D. Hymes (ed.). *The Use of Computers in Anthropology* (pp. 345-356). The Hague: Mouton & CO.

Parente, B., Queiroga, H., Teixeira, E., Sotto-Mayor, R., Barata, F., Sousa, A., ... e Araújo, A. (2007). Estudo epidemiológico do cancro do pulmão em Portugal nos anos de 2000/2002. Revista Portuguesa de Pneumologia, 13(2), 255-265.

Park, H. M. (2009). *Regression Models for Binary Dependent Variables Using Stata, SAS, R, LIMDEP, and SPSS*. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.

Pestana, D. D., e Velosa, S. F. (2008). *Introdução à Probabilidade e à Estatística* (3ª ed.). Lisboa: Fundação Calouste Gulbenkian.

Phillips-Wren, G., Sharkey, P., e Dy, S. M. (2008). Mining lung cancer patient data to assess healthcare resource utilization. Expert Systems with Applications, 35(4), 1611-1619.

Pinker, S. (1997). *How the Mind Works*. London. UK: The Penguin Press.

Piramuthu, S. (2004). Evaluating feature selection methods for learning in data mining applications. *European journal of operational research, 156*(2), 483-494.

PORDATA. (2015). Precipitação total. Consultado a 30 de Setembro de 2015 em http://www.pordata.pt/Portugal/Ambiente+de+Consulta/Tabela

QGIS. (s.d.). Home. Consultado a 12 de Fevereiro de 2015 em http://www.qgis.org/en/site/index.html

Rahm, E. (2000). Data Cleaning: Problems and Current Approaches. *Informatica*, 1-11.

Razali, N. M., e Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33.

Registo Oncológico Nacional 2006. (2012). Instituto Português de Oncologia de Lisboa de Francisco Gentil – EPE. Lisboa.

Registo Oncológico Regional Sul. (2014). Incidência, Sobrevivência e Mortalidade por cancro na região sul de Portugal – ISM 2008 | 2009. Registo Oncológico Regional Sul, Lisboa.

Registo Oncológico Regional Sul. (s.d.). Sobre nós – Quem Somos. Consultado a 24 de Julho de 2015 em http://www.ror-sul.org.pt/Apresentacao/Pages/QuemSomos.aspx

Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., ... e Brunak, S. (2011). Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol, 7(8), e1002141.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, 53-65.

The R Foundation. (s.d.). What is R? Consultado a 14 de Setembro de 2015 em https://cran.r-project.org/

The Univeristy of Waikato. (s.d.). Software. Consultado a 14 de Setembro de 2015 em http://www.cs.waikato.ac.nz/ml/weka/index.html

Vidyattama, Y. (2014). Issues in applying spatial autocorrelation on Indonesia's provincial income growth analysis.

Weed, L. L. (1968). Special article: Medical records that guide and teach. *New England Journal of Medicine, 278*(12), 593-600.

World Health Organization. (2015). Cancer [Fact Sheet Nº297]. Consultado a 6 de Agosto de 2015 de http://www.who.int/mediacentre/factsheets/fs297/en/

Yeh, D. Y., Cheng, C. H., e Chen, Y. W. (2011). A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications, 38*(7), 8970-8977.

Zhao, Y., e Karypis, G. (2002, November). Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 515-524). ACM.

Zhou, X., Chen, S., Liu, B., Zhang, R., Wang, Y., Li, P., ... e Yan, X. (2010). Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artificial Intelligence in Medicine, 48(2), 139-152.

APÊNDICES

A - CÓDIGO R DA TAREFA DE INFERÊNCIA ESTATÍSTICA

```
##############################################
###     Comparação de Valores Médios       ###
##############################################
# Q-Q plot
#sexo Masculino
qqnorm(rorM$IdadeDiag,main=NULL,xlab="Quantil Teórico",ylab="Quantil Amostral")
qqline(rorM$IdadeDiag,col=2,lwd=2)
#sexo Feminino
qqnorm(rorF$IdadeDiag,main=NULL,xlab="Quantil Teórico",ylab="Quantil Amostral")
qqline(rorF$IdadeDiag,col=2,lwd=2)

## Teste Shapiro-W.
shapiro.test(rorM$IdadeDiag)
shapiro.test(rorF$IdadeDiag)

## Teste Kolmogorov-S.
ks.test(rorM$IdadeDiag,rorF$IdadeDiag)

##############################################
###         Tabelas de Contingência        ###
##############################################
teste.indep<-function(tabela){
# FUNÇÃO PARA O TESTE DE INDEPENDÊNCIA EM TABELAS DE CONTINGÊNCIA
# argumento de entrada - tabela de contagens (n*p)
  n<-dim(tabela)[1]; p<-dim(tabela)[2]
  test<-chisq.test(tabela,simulate.p.value=TRUE,B=5000)
  fe<-matrix(test$expected,n,p)  # obter a tabela de freq. esperadas
  x<-as.integer((n*p)*.8)
  if((sum(fe>=1)==(n*p)) & (sum(fe>=5)>=x)){  # restrições do teste Qui-Quadrado
        test.name<-"Teste de Independência do Qui-Quadrado"
        p<-test$p.value
      }else{  # teste de fisher
        test.name<-"Teste Exato de Fisher"
        p<-fisher.test(tabela,B=5000)$p.value
     }
      alpha<-0.05
        if(p<alpha) {decisao<-"Rejeitamos H0"} else {decisao<-"Não rejeitamos H0"}
        lista<-list(tabela,round(fe,2),test.name,p,decisao)
        names(lista)<-c("freq.observada","freq.esperada","teste","p.value","resultado")
        return(lista)
}

########################################
##             REGRESSÃO              ##
########################################
# CONCELHOS - regressão múltipla
# TIBMF ~ IQArMelhor + IdadeMeanDiag.porConcelho
t<-glm(TIBMF~IQArMelhor+vMeanIdade,data=dados)
summary(t)

# PACIENTES -> Y - estadio: logit binary
ror2<-ror[Estadio=="Doença local ou locoregional" | Estadio=="Doença metastática",]
ror2$EstadioBinary<-vector(mode="integer",length=dim(ror2)[1])
ror2$EstadioBinary[Estadio=="Doença metastática"]<-1

# Estadio ~ IdadeDiag + IQAr + Idade*IQAr
t<-
glm(EstadioBinary~IdadeDiag+IQAr+IdadeDiag*IQAr,family=binomial(link="logit"),data=ror2)
summary(t)
```

B - CÓDIGO R DA ANÁLISE DE DADOS ESPACIAIS

```
#######################################
##              Funções              ##
#######################################
getw<-function(mapa){
# Retorna uma lista de pesos de um mapa
  vizinhos<-poly2nb(mapa,queen=TRUE)  # criar lista de vizinhos para cada região
  pesos<-nb2listw(vizinhos,zero.policy=TRUE)  # lista de pesos, por região, padronizadas
por linha
  return(pesos)
}

getglobal<-function(var,pesos){
# Retorna o teste de moran para uma variável
  p1<-moran.test(var,pesos,alternative="two.sided",zero.policy=TRUE)$p.value  # teste I
de Moran
  p2<-moran.mc(var,pesos,999,zero.policy=TRUE)$p.value  # permutação do Teste I de Moran
  alpha<-.05;
  if(p1>alpha & p2>alpha) {tmoran<-1}  # 1 - nao rejeitar H0
  else {if(p1<alpha & p2<alpha) {tmoran<-2}  # 2 - rejeitar H0
  else tmoran<-3} # 3 - testes em desacordo
  # 0 - não foi realizado o teste
  return(tmoran)
}

getlocal<-function(mapa,var,var.nome,pesos){
# Associação Espacial Local
  G<-localG(var,pesos,zero.policy=TRUE)  # teste G local
  lisa<-localmoran(var,pesos,zero.policy=TRUE)  # teste LISA
  I<-lisa[,1]
  quadrante<-vector(mode="numeric",length=dim(lisa)[1])
  names(quadrante)<-mapa$Municipio
  alpha<-.05
  I<-I-mean(I)  # centralizar para se usar o g.cartesiano
  var<-var-mean(var)
  quadrante[var > 0 & I > 0 & G > 0]<-4  # alto-alto 1°Q
  quadrante[var < 0 & I < 0 & G < 0]<-1  # baixo-baixo 3°Q
  quadrante[var < 0 & I > 0 ]<-2  # baixo-alto 2°Q
  quadrante[var > 0 & I < 0]<-3  # alto-baixo 4°Q
  quadrante[lisa[,5] > alpha]<-0
  q<-c(0,1,2,3,4)
  cores<-c("white","blue",rgb(0,0,1,alpha=0.4),rgb(1,0,0,alpha=0.4),"red")
  png(paste(var.nome,".png",sep=""),width = 5*300,height = 5*300,res = 330,pointsize=11)

plot(mapa,border="gray51",col=cores[findInterval(quadrante,q,all.inside=FALSE)],sub=var.
nome)
  box()
  legend("bottomright",title="Classe",legend=c("insignificante","baixo-baixo","baixo-
alto","alto-baixo","alto-alto"),fill=cores,cex=.7)
  dev.off()
  valorsem<-names(quadrante)[quadrante == 1 | quadrante == 4]  # concelhos semelhantes
  valordsem<-names(quadrante)[quadrante == 2 | quadrante == 3]  # concelhos
dissemelhantes
  return(list(vs=valorsem,vds=valordsem))
}

getresume<-function(mapa,v,nome){
# Resumo das frequências absolutas
  aux<-as.vector(table(v))  # vetor de semelhança ou dissemelhança
  names(aux)<-names(table(v))
  laux<-length(aux)  #número de linhas
  lfabs<-dim(mapa)[1]  # construção do vector de freq. abs.
  fabs<-vector(mode="numeric",length=lfabs)  #número de linhas
  names(fabs)<-mapa$Municipio
  for(i in 1:lfabs){
    for(k in 1:laux){
      if(names(fabs)[i]==names(aux)[k]) fabs[i]<-aux[k]
    }
  }
  mapa$freqAbsolutas<-fabs
  cor<-sample(c("YlOrBr","YlGnBu","YlGn","BuPu"),1)
```

```
  imagem<-
spplot(mapa,"freqAbsolutas",col.regions=colorRampPalette(brewer.pal(9,cor))(18))
  png(paste(nome,".png",sep=""),width=5*300,height=5*300,res=330,pointsize=11)
  print(imagem)
  dev.off()
}

analysis<-function(mapa,indice){
# Análise de Dados Espaciais
  p<-dim(mapa)[2]  # número de variáveis total
  pesos<-getw(mapa)  # pesos espaciais
  tmoran<-vector(mode="numeric",length=(dim(mapa)[2]-(indice-1)))  # MORAN I
  names(tmoran)<-names(mapa)[indice:p]
  # armazenar concelhos
  semelhante<-c()
  dissemelhante<-c()
  # autocorrelação e associação espacial para cada variável de interesse
  j<-1
  for(i in indice:p){
    var<-mapa@data[,i]
    var.nome<-names(mapa)[i]
    if(sum(var)!=0){  # a análise não é realizada para vetores só com zeros
      tmoran[j]<-getglobal(var,pesos)  # teste I de Moran
      if(tmoran[j]==1){ # se não rej. H0 então podemos avançar para a análise local
        l1<-getlocal(mapa,var,var.nome,pesos)  # associação espacial local
        semelhante<-c(semelhante,l1$vs)  # armazenar concelhos com valores semelhantes
        dissemelhante<-c(dissemelhante,l1$vds)  # armazenar concelhos com valores
dissemelhantes
      }
    }
    j<-j+1
  }
  # imagem com as freq abs dos concelhos sem. e disse. de todas as var de interesse
  getresume(mapa,semelhante,"ResumoSemelhança")
  getresume(mapa,dissemelhante,"ResumoDissemelhança")
  # Criar a lista de informações sobre o Teste de Moran
  explic<-c("1 - nao rejeitar H0 ","2 - rejeitar H0 ","3 - testes em desacordo","0 -
teste não realizado")
  lista<-list(explic,table(tmoran),prop.table(table(tmoran)),tmoran)
  names(lista)<-c("Explicação dos Números","Tabela Teste de Moran","Tabela Proporção
Teste de Moran","Teste de Moran Individual")
  ficheiro<-file("Info_Moran.csv", open="a")  # escrever o ficheiro de output dos dados
  t<-length(lista)
  for (i in 1:t){

write.table(names(lista)[i],file=ficheiro,sep=";",quote=FALSE,col.names=FALSE,row.names=
FALSE)
    write.table(lista[[i]],
file=ficheiro,sep=";",quote=FALSE,col.names=FALSE,row.names=TRUE)
  }
  close(ficheiro)
}
```

C – CÓDIGO R DA ANÁLISE DE AGRUPAMENTOS

```
########################################
##              Funções               ##
########################################
hierarquico<-function(md,nomeTeste,l){
# Função para implementar o método de agrupamento hierarquico
# aglomerativo segundo uma matriz de distância e para todos os metodos em estudo
  metodos<-c("average","single","complete","ward.D2")
  m<-length(metodos)
  alturas.valor<-c(1,.8,.6,.4,.3)  # altura de corte: maxTreeHeight
  a<-length(alturas.valor)
  alturas<-paste("Altura = ",alturas.valor,sep="")
  matrizk<-matrix(0,nrow=a,ncol=m)  # k - numero de grupos
  coefA<-vector(mode="numeric",length=m)  # coeficiente do agrupamento
  for(j in 1:m){
    for(i in 1:a){
      h1<-hclust(md,method=metodos[j])
      h1$height<-sort(h1$height)
      numerk<-
cutreeDynamicTree(h1,maxTreeHeight=alturas.valor[i],deepSplit=TRUE,minModuleSize=25)
      matrizk[i,j]<-max(numerk)
      coefA[j]<-coef.hclust(h1)
    } }
  names(coefA)<-colnames(matrizk)<-metodos; rownames(matrizk)<-alturas
  return(list(matrizk,coefA))
}

dist.combinacao<-function(dados,variaveis,c,l){
# Função para construir a matriz de distancias combinada
# Mahalanobis (atributos continuos) e Hamming (atributos categoricos)
  p<-length(variaveis); vfactor<-c(); vnumero<-c()
  for(i in 1:c){
    if(i %in% variaveis) {if(class(dados[,i])=="factor"){
      vfactor<-c(vfactor,i) } else vnumero<-c(vnumero,i)
    }
  }
  # Matriz distancias - Mahalanobis
  d1<-mahalanobis.dist(dados[,vnumero],vc=NULL)  # matriz das covariancias
  matrizM<-as.matrix(d1);pc<-length(vnumero)
  # Matriz semelhança - Hamming
  pd<-length(vfactor)
  matrizH<-md<-matrix(nrow=l,ncol=l)  # md = matriz de distancias combinada
  for(i in 1:l){
    for(j in seq(i,l)){
      matrizH[j,i]<-matrizH[i,j]<-(sum(dados[i,vfactor]==dados[j,vfactor],na.rm=T)/pd)
      md[j,i]<-md[i,j]<-round(((pc/p)*matrizM[i,j])+((pd/p)*matrizH[i,j]),2)
    }
  }
  return(as.dist(md))
}

agrupamento<-function(dados,variaveis,medida,nomeTeste,func){
# Função conjunta dos métodos de agrupamento #
# argumentos de entrada:
    # conjunto de dados | variaveis (vector numerico) | medida de semelhança|
    # metodo de agrupamento | func para apresentacao dos resultados
  # PARTE I
  l<-dim(dados)[1]
  c<-dim(dados)[2]
  # AGRUPAMENTO HIERARQUICO DOS DADOS
  # Obj: obter numero de grupos e os coeficientes dos agrupamentos
  ## Medida de Semelhança GOWER
  if(medida=="gower"){
    md<-daisy(dados[,variaveis],stand=T,metric="gower") #stand=padronizar
  }
  ## Medida: combinação linear de Mahalanobis (continuos) e Hamming (categoricos)
  else{
    md<-dist.combinacao(dados,variaveis,c,l)
  }
  ## Agrupamento Hierarquico
  h<-hierarquico(md,nomeTeste,l)
```

```
  matrizks<-h[[1]]
  coefsA<-h[[2]]
  ks<-as.vector(matrizks); ks<-unique(ks)  # NUMERO DE GRUPOS UNICOS
  ks<-ks[!is.na(ks)]  # remover NA se existir
  quantosk<-length(ks); remover<-c()
  for(i in 1:quantosk){  # k: remover o valor zero e um
    if(ks[i]==0|ks[i]==1){ remover<-c(remover,i) }  # armazenar os indices
  }
  if(length(remover)>=1){  # remover os indices encontrados
    ks<-ks[-remover]; quantosk<-length(ks)
  }
  # PARTE II
  # PARTICIONAMENTO DOS DADOS
  # Obj: construção do particionamento dos dados segundo os k's obtidos
  resultado<-matrix(nrow=l,ncol=quantosk)
  colnames(resultado)<-paste("Part",ks,sep="")
  listaInfo<-vector(mode="list",length=quantosk)
  for(i in 1:quantosk){
    part<-pam(md,ks[i],diss=TRUE)
    cmd<-cmdscale(md)  # escala multidimensional
    grupos <- levels(factor(part$clustering))
    # representação gráfica
    jpeg(filename=paste(nomeTeste,"_PLOT.DIST_",ks[i],".jpeg",sep=""),width=5*300,
         height=5*300,res=330,pointsize=11)
    ordiplot(cmd,xlab="Dimensão1",ylab="Dimensão2")
    cores<-sample(c(8,12,30:31,32:33,36:37,41:42,51:53,75:76,
88:90,100:103,111:114,131:132,139,259,375,459,614,423,652),length(grupos))
    for(i in seq_along(grupos)){
        points(cmd[factor(a[,962])==grupos[i],],col=cores[i],pch=19)
    }
    dev.off()
    # armazenar info sobre a silhueta
    si<-silhouette(part$clustering,md)
    imagem<-fviz_silhouette(si,label=FALSE,print.summary=FALSE)
    png(paste(nomeTeste,"_PLOT.SILHUETA_",ks[i],".jpeg",sep=""))
    print(imagem)
    dev.off()
    # report
    resultado[,i]<-part$clustering
    listaInfo[[i]]<-part$clusinfo
  }
  # PARTE III
  # APRESENTACAO DOS RESULTADOS
  func(dados,nomeTeste,matrizks,coefsA,quantosk,ks,c,listaInfo,md,resultado)
  }
}
```

D – TAXAS DE INCIDÊNCIA BRUTA GLOBAL E POR SEXO

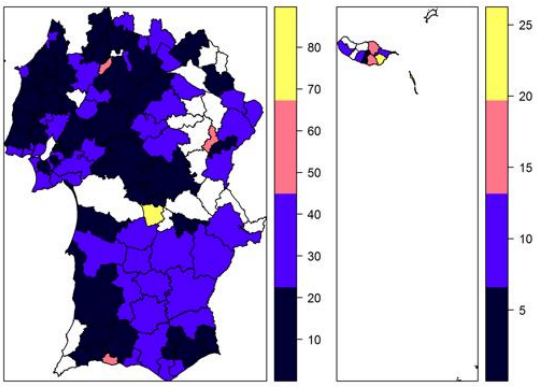

Figura D1 - Taxa de incidência bruta, por cada concelho abrangido pelo ROR-Sul, em 2013 e por 100 000 habitantes

Nota: as localizações em branco não têm registo de incidência

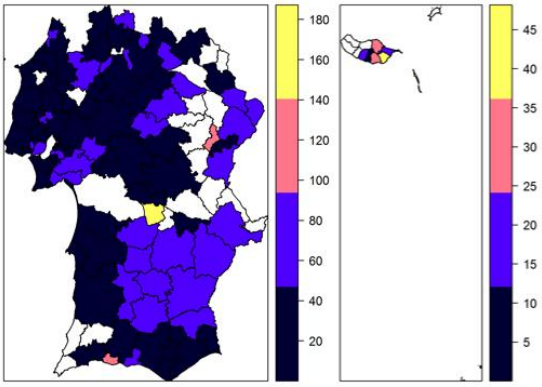

Figura D2 - Taxa de incidência bruta do sexo masculino, por cada concelho abrangido pelo ROR-Sul, em 2013 e por 100 000 habitantes

Nota: as localizações em branco não têm registo de incidência

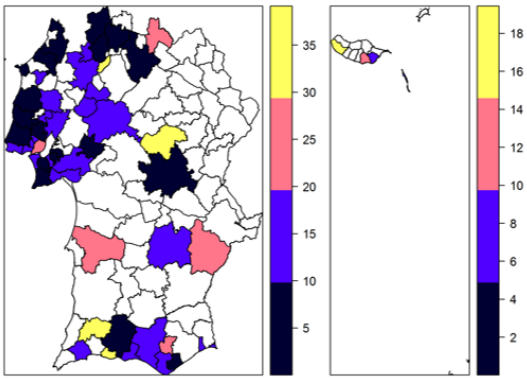

Figura D3 - Taxa de incidência bruta do sexo feminino, por cada concelho abrangido pelo ROR-Sul, em 2013 e por 100 000 habitantes

Nota: as localizações em branco não têm registo de incidência

E – FATORES AMBIENTAIS

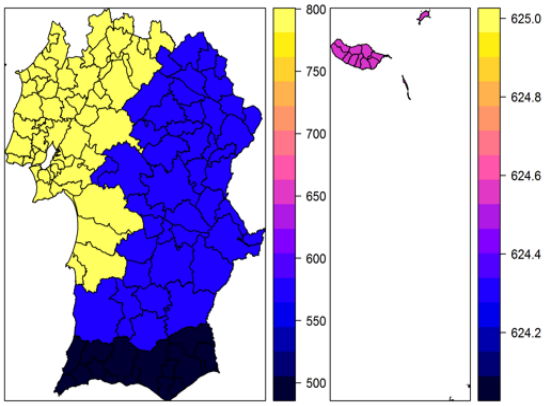

Figura E1 - Média anual de precipitação (mm), por cada região abrangida pelo ROR-Sul

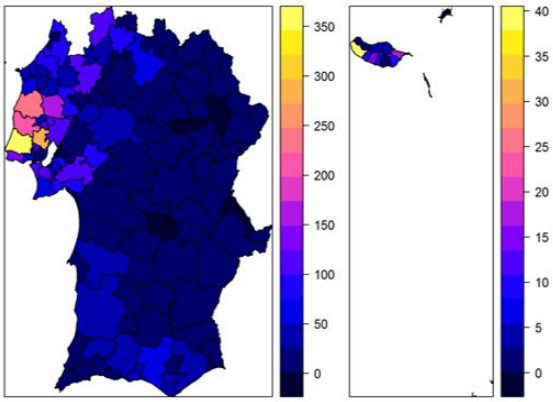

Figura E2 - Número médio anual de incêndios, por cada concelho abrangido pelo ROR-Sul

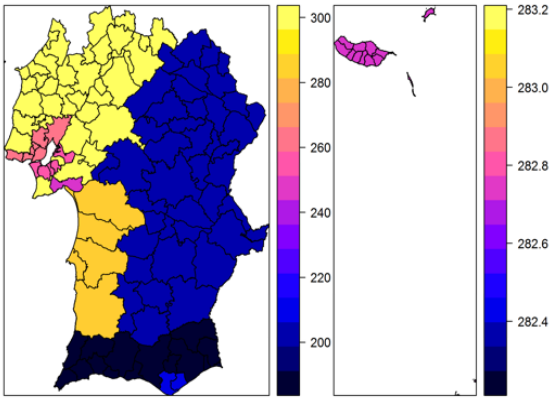

Figura E3 - Número médio de dias anuais da categoria muito bom e bom do IQA, por cada zona abrangida pelo ROR-Sul

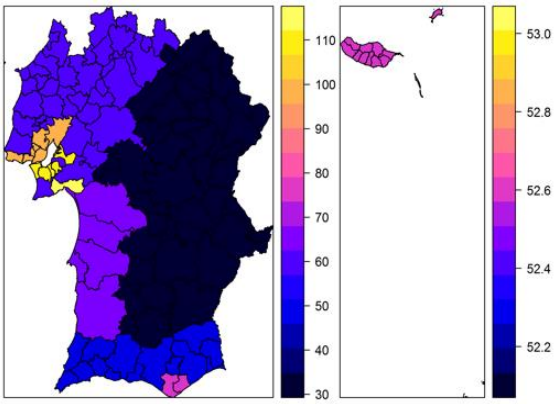

Figura E4 - Número médio de dias anuais da categoria médio, fraco e mau do IQA, por zona abrangida pelo ROR-Sul

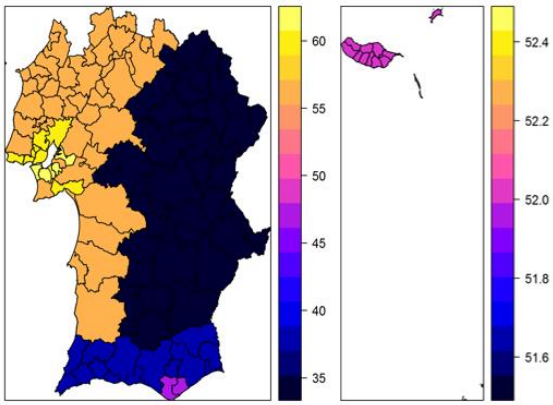

Figura E5 - Indicador da qualidade do ar por cada zona abrangida pelo ROR-Sul

F – GRÁFICOS Q-Q DA IDADE DOS PACIENTES POR SEXO

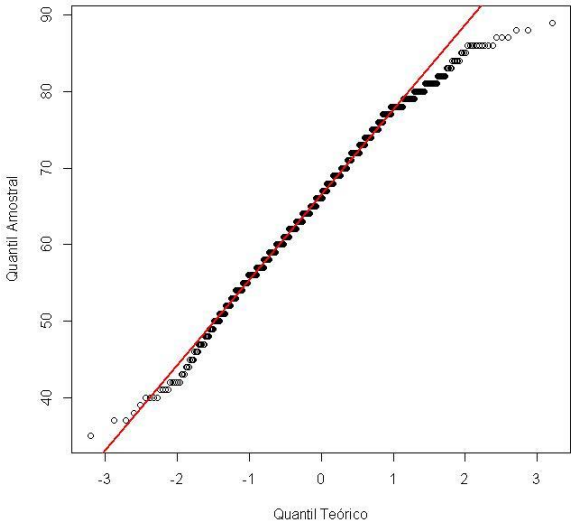

Figura F1 – Gráfico Q-Q da idade dos pacientes do sexo masculino

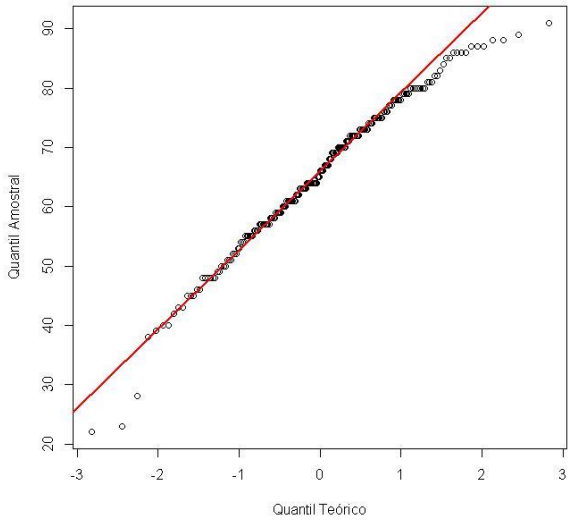

Figura F2 – Gráfico Q-Q da idade dos pacientes do sexo feminino

G – RESULTADOS DA REGRESSÃO MÚLTIPLA GERAIS

Tabela G1 – Principais resultados obtidos da regressão múltipla (variável-resposta: *Tx Incidência*)

Modelo 1	AIC = 976,39	R^2 = 0,4085
Variáveis	**Coeficientes das Variáveis**	**p-value do Teste *T-Student***
MedIdade	0,3731	1,16E-15
IQAr	-0,27003	0,0103
Modelo 2	AIC = 977,01	R^2 = 0,4149
Variáveis	**Coeficientes das Variáveis**	**p-value do Teste *T-Student***
MedIdade	0,367826	2,76E-15
IQArMelhor	-0,0648	0,01311
IQArPior	0,002413	0,96751
Modelo 3	AIC = 975, 01	R^2 = 0,4149
Variáveis	**Coeficientes das Variáveis**	**p-value do Teste *T-Student***
MedIdade	0,36811	9,42E-16
IQArMelhor	-0,06429	0,00487
Modelo 4	AIC = 981,39	R^2 = 0,3845
Variáveis	**Coeficientes das Variáveis**	**p-value do Teste *T-Student***
MedIdade	0,36217	1,45E-14
IQArPior	-0,06969	0,189

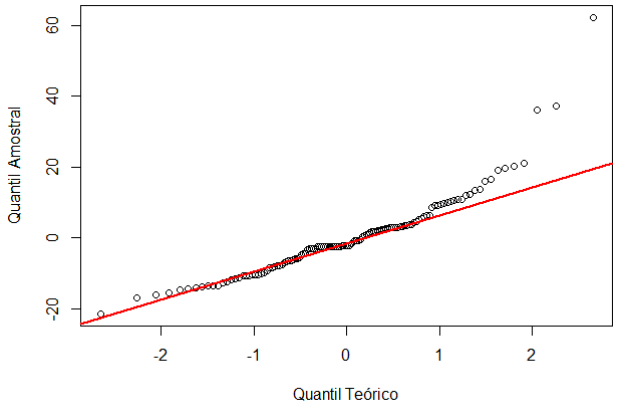

Figura G1 – Gráfico Q-Q Normal dos resíduos do modelo 3 (Tabela G1)

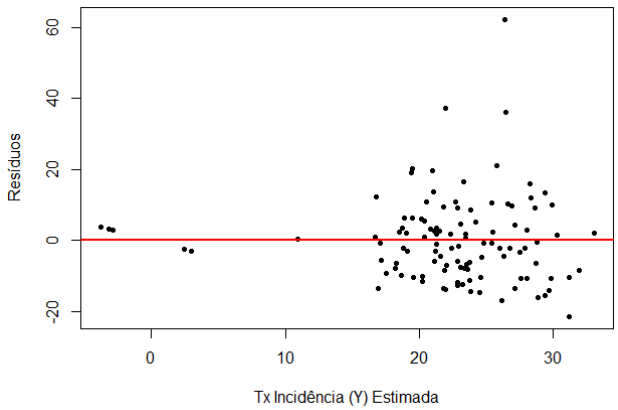

Figura G2 – Observações estimadas e resíduos correspondentes

H – DENDROGRAMA (MÉTODO *WARD*)

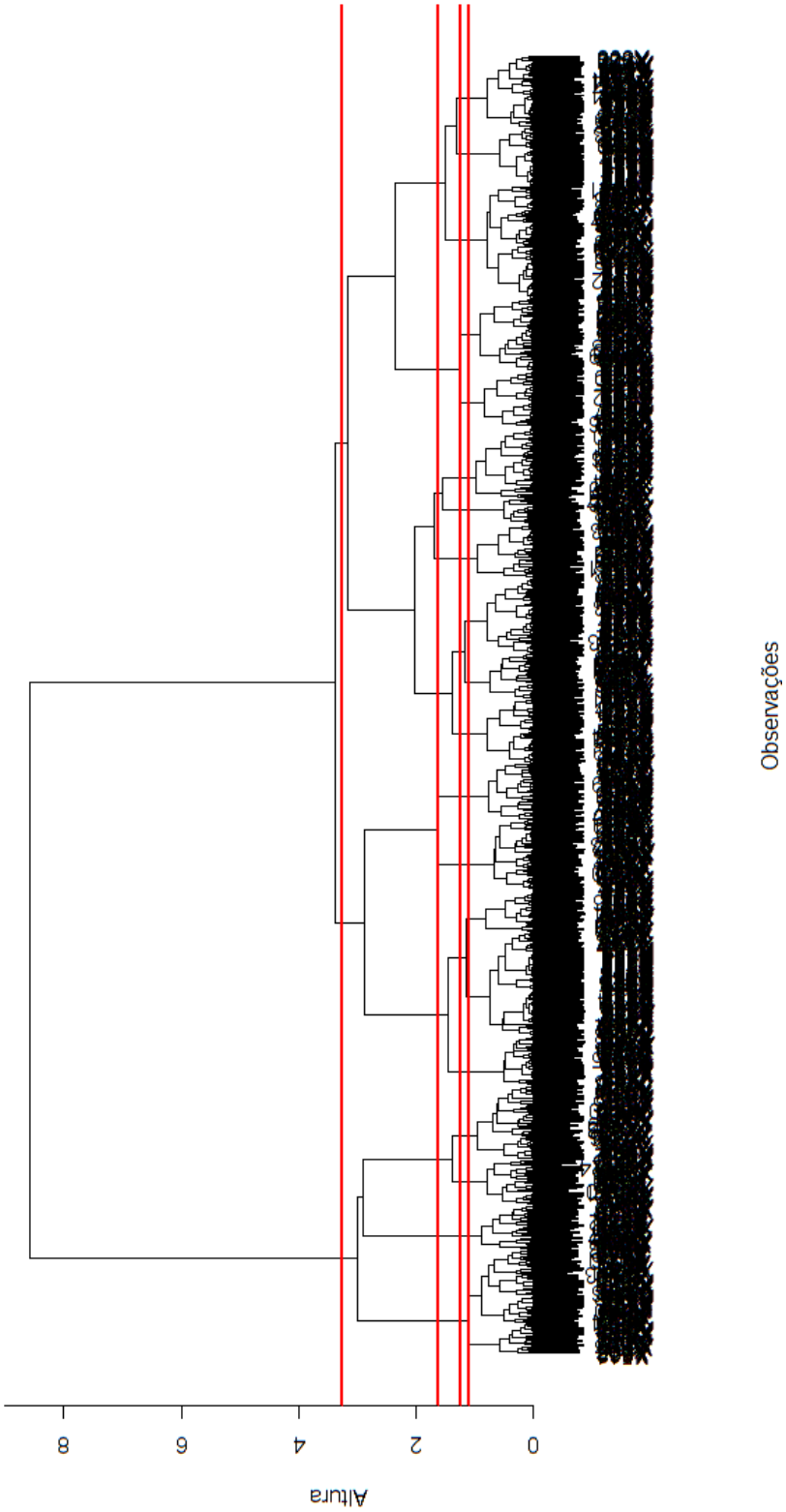

Figura H1 – Dendrograma obtido através do método ***ward***

I – Representação Gráfica do Particionamento de Dados em 10 Grupos

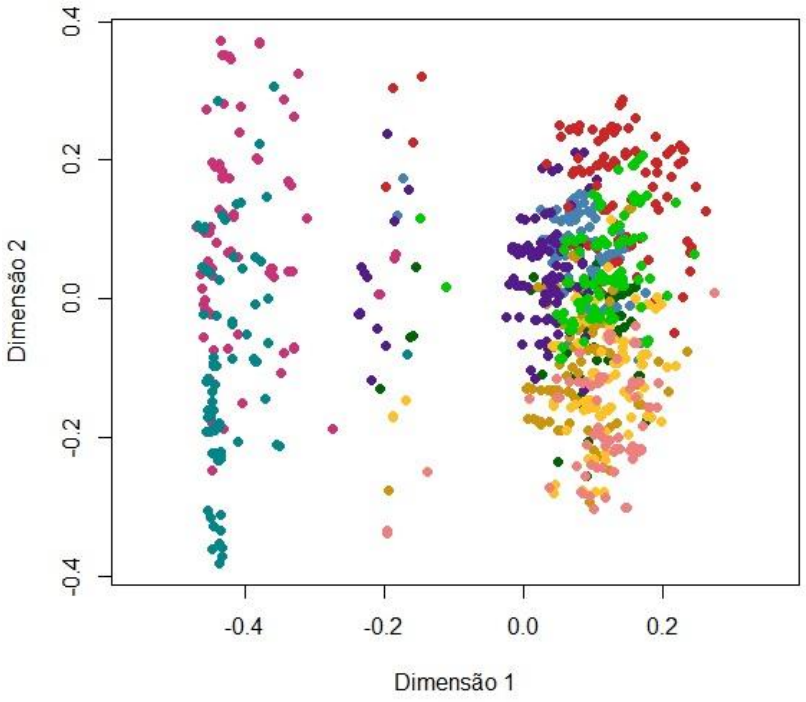

Figura I1 – Representação das observações considerando 10 grupos

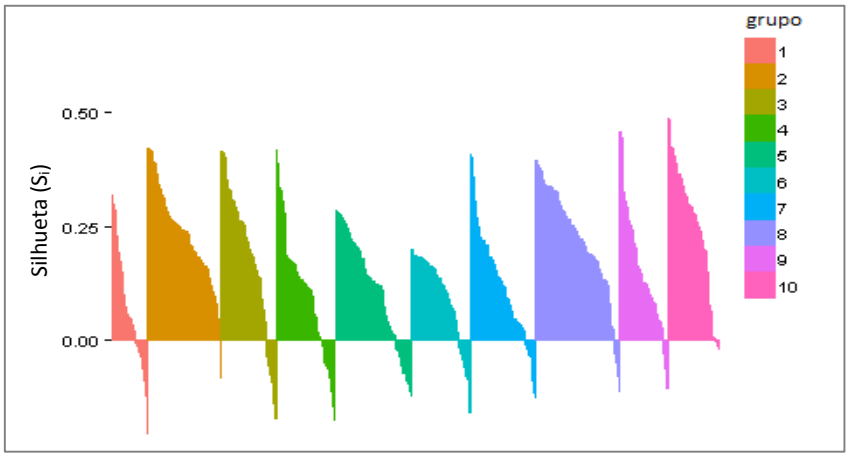

Figura I2 - Informação da silhueta das observações considerando 10 grupos

ANEXOS

ANEXO 1 - REGIÃO ROR-SUL EM DETALHE

Tabela 1 - Legenda da Figura 1

ID	Concelho	Distrito/Ilha	ID	Concelho	Distrito
0	CÂMARA DE LOBOS	ILHA DA MADEIRA	147	ALENQUER	LISBOA
1	SANTA CRUZ		148	ARRUDA DOS VINHOS	
2	PORTO SANTO	ILHA DE PORTO SANTO	149	AZAMBUJA	
3	PORTO MONIZ	ILHA DA MADEIRA	150	CADAVAL	
4	SÃO VICENTE		151	CASCAIS	
5	SANTANA		152	LISBOA	
6	PONTA DO SOL		153	LOURES	
7	FUNCHAL		154	LOURINHÃ	
8	CALHETA		155	MAFRA	
9	RIBEIRA BRAVA		156	OEIRAS	
10	MACHICO		157	SINTRA	
19	ALJUSTREL	BEJA	158	SOBRAL DE MONTE AGRAÇO	
20	ALMODÔVAR		159	TORRES VEDRAS	
21	ALVITO		160	VILA FRANCA DE XIRA	
22	BARRANCOS		161	AMADORA	
23	BEJA		162	ODIVELAS	
24	CASTRO VERDE		163	ALTER DO CHÃO	PORTALEGRE
25	CUBA		164	ARRONCHES	
26	FERREIRA DO ALENTEJO		165	AVIS	
27	MÉRTOLA		166	CAMPO MAIOR	
28	MOURA		167	CASTELO DE VIDE	
29	ODEMIRA		168	CRATO	
30	OURIQUE		169	ELVAS	
31	SERPA		170	FRONTEIRA	
32	VIDIGUEIRA		171	GAVIÃO	
87	ALANDROAL	ÉVORA	172	MARVÃO	
88	ARRAIOLOS		173	MONFORTE	
89	BORBA		174	NISA	
90	ESTREMOZ		175	PONTE DE SOR	
91	ÉVORA		176	PORTALEGRE	
92	MONTEMOR-O-NOVO		177	SOUSEL	
93	MORA		196	ABRANTES	SANTARÉM

94	MOURÃO	
95	PORTEL	
96	REDONDO	
97	REGUENGOS DE MONSARAZ	
98	VENDAS NOVAS	
99	VIANA DO ALENTEJO	
100	VILA VIÇOSA	
101	ALBUFEIRA	FARO
102	ALCOUTIM	
103	ALJEZUR	
104	CASTRO MARIM	
105	FARO	
106	LAGOA	
107	LAGOS	
108	LOULÉ	
109	MONCHIQUE	
110	OLHÃO	
111	PORTIMÃO	
112	SÃO BRÁS DE ALPORTEL	
113	SILVES	
114	TAVIRA	
115	VILA DO BISPO	
116	VILA REAL DE SANTO ANTÓNIO	
131	ALCOBAÇA	LEIRIA
135	BOMBARRAL	
136	CALDAS DA RAINHA	
141	NAZARÉ	
142	ÓBIDOS	
144	PENICHE	

197	ALCANENA	
198	ALMEIRIM	
199	ALPIARÇA	
200	BENAVENTE	
201	CARTAXO	
202	CHAMUSCA	
203	CONSTÂNCIA	
204	CORUCHE	
205	ENTRONCAMENTO	
206	FERREIRA DO ZÊZERE	
207	GOLEGÃ	
208	MAÇÃO	
209	RIO MAIOR	
210	SALVATERRA DE MAGOS	
211	SANTARÉM	
212	SARDOAL	
213	TOMAR	
214	TORRES NOVAS	
215	VILA NOVA DA BARQUINHA	
216	OURÉM	
217	ALCÁCER DO SAL	SETÚBAL
218	ALCOCHETE	
219	ALMADA	
220	BARREIRO	
221	GRÂNDOLA	
222	MOITA	
223	MONTIJO	
224	PALMELA	
225	SANTIAGO DO CACÉM	
226	SEIXAL	
227	SESIMBRA	
228	SETÚBAL	
229	SINES	

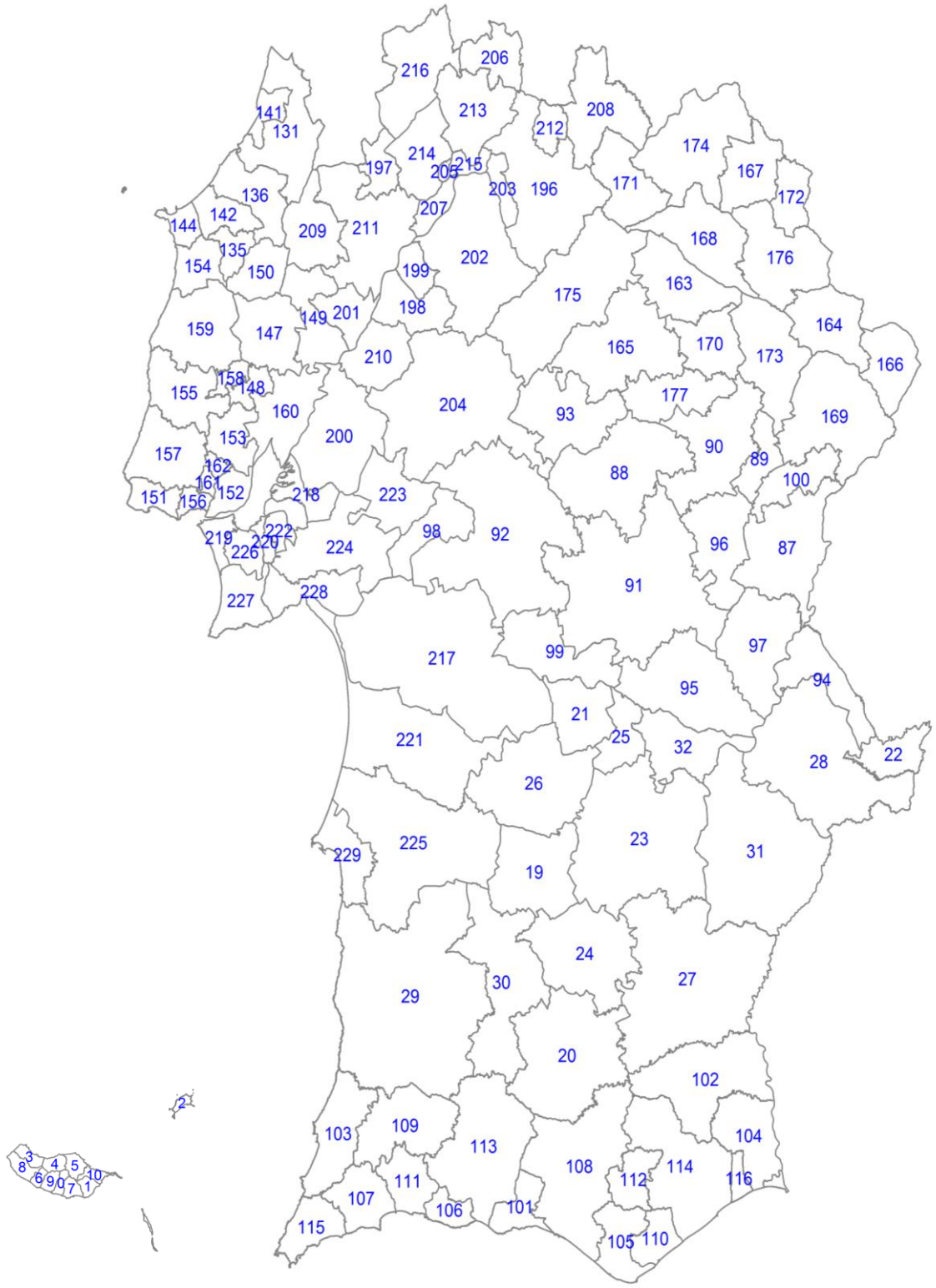

Figura 1 – Região ROR-Sul em detalhe

ANEXO 2 - VARIÁVEIS DO ROR-SUL DISPONÍVEIS PARA O ESTUDO

DATA COLLECTION BY USING TXT DATABASE
Version 2.3, August 10, 2014

Data common to all tumours included in HR studies

(in red variables common to the EUROCARE survival study)

	Variable	*Recommended coding of values*	*Format**	*Definition/notes*
1.	Registry	registry name	A10	Missing not allowed Please, insert as many "_" as necessary to reach the accepted 10-character string
2.	Identification code	Necessary for possible corrections and later updating. Unique id code for both HR and low resolution study (to update life status through record linkage with EUROCARE)	A20	Missing not allowed Please, insert as many "_" as necessary to reach the accepted 20-character string
3.	Tumour identification code	If available, in addition to patient's id code	A20	Please, insert as many "_" as necessary to reach the accepted 20-character string
4.	Gender	1= male; 2= female	F1	Missing not allowed
5.	Day of birth	01-31	F2	Differently from ENCR or EUROCARE rules, the HR study collects the exact date of birth, to compute exactly the age at diagnosis Missing not allowed
6.	Month of birth	01-12	F2	
7.	Year of birth	1900-1998	F4	
8.	Day of incidence	01-31	F2	Differently from ENCR or EUROCARE rules, the HR study collects the exact date of birth, to compute outcome indicators such as short term survival, operative mortality, disease-free interval. Missing not allowed
9.	Month of incidence	01-12	F2	
10.	Year of incidence	2011-2013 (possible range)	F4	
11.	Primary tumour site	Topography coded according to the international classification of diseases for oncology (ICD-O-3) C00.0-C80.9	A5	See the ENCR recommendations (e.g. C50.9) Missing not allowed
12.	Morphology (ICD-O)	Morphology coded according to the international classification of diseases for oncology (ICD-O-3) 8000-9989	A4	See the ENCR recommendations (e.g. 8000) Missing not allowed Note: in the HR study on Lymphoma, only the following morphological codes are admitted: Diffuse Large B-cell lymphoma (9678-9680, 9684); Follicular lymphoma (9690-9691, 9695, 9698).
13.	Behaviour	It corresponds to the 5th digit of ICD-O-3 code 2= in situ (available only for breast) 3= malignant	F1	Benign tumours not to be included Missing not allowed Note: Please, for Lymphomas record the following code: 3

14.	Grading	It corresponds to the 6th digit of ICD-O-3 code 1= Grade I, well differentiated; 2= Grade II, moderately differentiated; 3= Grade III, poorly differentiated; 4= Grade IV, undifferentiated; 5= T-cell; 6= B-cell; 7= Null cell; 8= NK cell; 9= Not determined, not graded;	F1	Note 1 for BREAST cancer: In case of two different gradings (one at biopsy and one at the first surgery without -or before- neo-adjuvant chemotherapy), the highest grade should be collected Note 2 for LYMPHOMAS: Codes 1-4 have not to be recorded. Please record 5 to 8; the 6th digit of the morphology code can be added to the cell types of lymphomas that have been identified by surface marker studies. Code any statement of T-cell or B-cell involvement whether or not marker studies are documented in the patient record.
15.	Basis of diagnosis	0 = DCO; 1= clinical (clinical only, clinical investigation, tumour marker) 2= microscopic (histology of a primary tumour, cytology, histology of a metastasis); 9 = unknown	F1	See the ENCR recommendations
16.	Incidental finding of cancer at the autopsy	1= yes; 2= no; 9= not allowed	F1	
17.	Multiple tumour	In case of multiple tumour, please report the temporal ranking of the cancer under study (=that for which HR variables are being collected), e.g. 1, 2, 3, etc. This number corresponds to that actually collected by the CR, regardless the inclusion criteria of this study	F1	For synchronous bilateral breast cancer, keep the ENCR rule: fill-in two records, although in the analyses only one case will be considered. Bilateral synchronous lung cancer will be considered as a single case.
18.	Inclusion in controlled clinical trial (optional)	1= yes; 2= no; 9= unknown	F1	
19.	Multidisciplinary team consulting meeting (optional)	1= yes; 2= no; 9= unknown	F1	"Yes" means that a multidisciplinary team meeting is documented in the clinical notes
	19 items for co-morbidities at diagnosis (Charlson index)			
20.	Charlson index (optional)	00-90 99=unkaown	F2	Record the total Charlson score if available in the clinical notes. However, score each item (see following diseases): See Annex C for the definition
21.	Myocardial infarct	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
22.	Congestive heart failure	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
23.	Peripheral vascular disease	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
24.	Cerebrovascular disease	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
25.	Dementia	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
26.	Chronic obstructive pulmonary disease	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
27.	Connective tissue disease/ Rheumatologic disease	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index

28.	Peptic ulcer disease	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
29.	Mild liver disease	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
30.	Diabetes without end organ damage	1= yes; 2= no; 9= unknown	F1	1: score for Charlson index
31.	Hemiplegia or Paraplegia	1= yes; 2= no; 9= unknown	F1	2: score for Charlson index
32.	Moderate to severe renal disease	1= yes; 2= no; 9= unknown	F1	2: score for Charlson index
33.	Diabetes with end organ damage	1= yes; 2= no; 9= unknown	F1	2: score for Charlson index
34.	Any malignant solid tumour (additional to that in study) without metastasis	1= yes; 2= no; 9= unknown	F1	2: score for Charlson index
35.	Lymphoma	1= yes; 2= no; 9= unknown	F1	2: score for Charlson index
36.	Leukemia	1= yes; 2= no; 9= unknown	F1	2: score for Charlson index
37.	Moderate or severe liver disease	1= yes; 2= no; 9= unknown	F1	3: score for Charlson index
38.	Metastatic solid tumour	1= yes; 2= no; 9= unknown	F1	6: score for Charlson index
39.	AIDS/HIV	1= yes; 2= no; 9= unknown	F1	6: score for Charlson index
40.	Smoker (optional)	1= yes, currently; 2= yes, previously; 3= no, never; 9= unknown	F1	
41.	BMI (optional)	00.0-60.9 99.9= unknown	F4	Note: BMI could be also referred to the date of treatment Please, report directly the BMI value calculated as ratio between weight in Kg and height in meters (Kg/m^2) (Ex: 15.9 or 45.0)
	Performance status			
42.	Type of scale (optional)	1= Karnofsky; 2= ECOG/WHO; 9= not available	F1	See Annex C for the definition
43.	Score of performance status (optional)	000-100 999= unknown	F3	Note: If ECOG or WHO scale are used, please record scores in the following way: 000, 001-005
	Follow-up *- in the short term (e.g. 1-year or at the end of first treatment)* *- long term follow-up will be updated in the future*			
44.	Relapse	For SOLID TUMOURS: 0= never 1= local; 2= as regional nodes or adjacent tissues/organs; 3= distant metastases; 9= unknown For LYMPHOMAS: 0= no relapse; 1= relapse; 9= unknown	F1	Note for Lymphomas: a relapse may only occur after a complete remission
45.	Day of relapse	01-31 If missing or not applicable, code: 99	F2	
46.	Month of relapse	01-12 If missing or not applicable, code: 99	F2	

47.	Year of relapse	2011-XXXX If missing or not applicable, code: 9999	F4	
48.	Second cancer	C00.0-C80.9 C99.9, if unavailable; 999.9, if absent;	A5	Subsequent tumour diagnosed after the tumour under study to be coded according to ICD-O-3 code, if available; (Ex: C50.9)
49.	Day of incidence of second cancer	01-31 If missing or not applicable, code: 99	F2	
50.	Month of incidence of second cancer	01-12 If missing or not applicable, code: 99	F2	
51.	Year of incidence of second cancer	2011-XXXX If missing or not applicable, code: 9999	F4	
52.	Life status at last known contact	1= alive; 2= dead	F1	Missing not allowed
53.	Cause of death	Coded according to ICD-10 code for cause of death A00.0 – Z99.9 if available; 999.9, if unavailable	A5	Note: ICD-10 covers years from 1999 to present
54.	Day of last known contact	01-31	F2	Missing not allowed
55.	Month of last known contact	01-12	F2	Missing not allowed
56.	Year of last known contact	2011-XXXX	F4	Missing not allowed

*F=numeric; A=alphanumeric

Additional data for LUNG CANCER (C34)

	Variable	*Recommended coding of values*	*Format**	*Definition/notes*
57.	Laterality	1 = left; 2 = right; 3 = bilateral; 9 = unknown	F1	
58.	EGFR mutation	1= present; 2= absent; 9= unknown	F1	To be collected for non-small cell cancer ONLY
	Diagnostic exams on the lung within 3 months (after or before) from diagnosis			
59.	Conventional thorax Imaging (radiography, stratigraphy, CT scan)	1= done; 2= not done; 9= unknown	F1	
60.	Spiral CT	1= done; 2= not done; 9= unknown	F1	
61.	Positron emission tomography (PET)	1= done; 2= not done; 9= unknown	F1	
62.	Magnetic resonance imaging (MRI)	1= done; 2= not done; 9= unknown	F1	
63.	Bronchoscopy	1= done; 2= not done; 9= unknown	F1	
64.	Mediastinoscopy	1= done; 2= not done; 9= unknown	F1	
65.	Endobronchial ultrasound guided bronchoscopy	1= done; 2= not done; 9= unknown	F1	
	Diagnostic exams for distant metastases within 3 months (after or before) from diagnosis			
66.	Liver imaging	1= done; 2= not done; 9= unknown	F1	Each exam should be considered
67.	Lung imaging	1= done; 2= not done; 9= unknown	F1	
68.	Brain imaging	1= done; 2= not done; 9= unknown	F1	
69.	Skeleton imaging	1= done; 2= not done; 9= unknown	F1	
	Stage at diagnosis			
70.	pT	Pathological T, according to TNM classification, 7th revision. Two-character accepted strings are: XX ; 0_ ; is; 1_; 1a; 1b; 2_; 2a ; 2b ; 3_; 4_; 99= Not available	A2	
71.	cT	Clinical T, according to TNM classification, 7th revision. Two-character accepted strings are: XX ; 0_ ; is; 1_;1a; 1b; 2_; 2a ; 2b ; 3_; 4_; 99= Not available	A2	
72.	Pathological tumour diameter	Tumour size in mm. 000-900 988= >900 999= unknown	F3	
73.	Clinical tumour diameter	Tumour size in mm. 000-900 988= >900 999= unknown	F3	
74.	pN	Pathological N, according to TNM classification, 7th revision. One-character accepted strings are: X; 0; 1; 2; 3; + (for N+NOS); 9= Not available	A1	Note 1: "+" is not allowed by the TNM 7th revision, but allowed in this HR study as some cancer registries can only access this information
75.	cN	Clinical N, according to TNM classification, 7th revision. One-character accepted strings are: X; 0; 1; 2; 3; + (for N+NOS); 9= Not available	A1	For "+", see Note 1 on pN stage
76.	pM	Pathological M, according to TNM classification, 7th revision. 0=M0; 1=M1; 9= Not available	F1	

77.	cM	Clinical M, according to TNM classification, 7th revision. Two-character accepted strings are: 0 ; 1 ; 1a; 1b; 99= Not available	A2	
78.	Site of metastasis (in clear, optional)		A10	Please, report the site of metastasis. Otherwise, insert as many "_" as necessary to reach the accepted 10- character string
		Treatment		
79.	SURGERY	0= not done; 1= lobectomy; 2= pneumectomy; 3= partial resection; 4= segmentectomy; 8= done, but unknown type of surgery; 9= unknown	F1	
80.	Day of surgery	01-31 If missing or not applicable, code: 99	F2	
81.	Month of surgery	01-12 If missing or not applicable, code: 99	F2	
82.	Year of surgery	2011-2014 If missing or not applicable, code: 9999	F4	
83.	Reasons for no surgery	1= medical contraindications; 2= patient refusal; 3= advanced cancer; 4= other; 8= no indication; 9= unknown	F1	
84.	Surgical radicality	1 = R0, no residual tumour; 2 = R1, microscopic residual tumour; 3 = R2, macroscopic residual tumour; 4 = R1/R2, presence of residual tumour but unknown if R1 or R2; 9 = RX, presence of residual tumour cannot be assessed or information is not available	F1	
85.	CHEMOTHERAPY	1= done; 2= not done; 9= unknown	F1	
86.	Day starting chemotherapy	01-31 If missing or not applicable, code: 99	F2	
87.	Month starting chemotherapy	01-12 If missing or not applicable, code: 99	F2	
88.	Year starting chemotherapy	2011-2014 If missing or not applicable, code: 9999	F4	
89.	Modality of chemotherapy	1= neo-adjuvant; 2= adjuvant; 3= palliative; 9= unknown	F1	
90.	Type of chemotherapy (in clear, optional)		A20	Please, report the therapeutic scheme or the drug(s) used. Otherwise, insert as many "_" as necessary to reach the accepted 20-character string
91.	TARGETED TREATMENT (e.g. monoclonal	1= done; 2= not done; 9= unknown	F1	

	antibody)			
92.	Day starting targeted treatment	01-31 If missing or not applicable, code: 99	F2	
93.	Month starting targeted treatment	01-12 If missing or not applicable, code: 99	F2	
94.	Year starting targeted treatment	2011-2014 If missing or not applicable, code: 9999	F4	
95.	Type of targeted treatment (in clear, optional)		A20	Please, report the drug(s) used. Otherwise, insert as many "_" as necessary to reach the accepted 20-character string
96.	RADIOTHERAPY	1= done; 2= not done; 9= unknown	F1	
97.	Modality of radiotherapy	1= neo-adjuvant; 2= adjuvant; 3= palliative; 9= unknown	F1	
98.	Day starting radiotherapy	01-31 If missing or not applicable, code: 99	F2	
99.	Month starting radiotherapy	01-12 If missing or not applicable, code: 99	F2	
100.	Year starting radiotherapy	2011-2014 If missing or not applicable, code: 9999	F4	
101.	Reasons for no radiotherapy	1= medical contraindications; 2= patient refusal; 3= other; 8= no indication; 9= unknown	F1	

*F=numeric; A=alphanumeric